PG473 - Endbericht

Bertram Bödeker
Jana Ehlers
Christian Friem
René Goebels
Christoph Hübinger
Ahmet Kara
Markus Matz
Niels Pothmann
Martin Prause
Stefan Rosas
Mehmet Sari
Madan Sathe

Betreuung:
Prof. Dr. Bernd Reusch
Stefan Berlik

8. Februar 2006

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

1.1 Thema der PG

Mit privaten und beruflichen Veränderungen ändern sich auch die finanziellen Rahmendaten im Leben eines jeden Anlegers. Die Bedeutung einer auf die jeweiligen Bedürfnisse des Anlegers ausgerichtete Investmentstrategie hat sich gerade im Hinblick auf den demographischen Wandel drastisch verändert. Berufseinsteiger haben z.B. im Hinblick auf Sicherheit, Flexibilität oder Liquidität andere Bedürfnisse als Personen im Ruhestand. Das Internet kann dem Anleger dabei heutzutage vielfältige Informationen aus der Finanzwelt in unterschiedlicher Qualität, Quantität, Aussagekraft und Zeitlichkeit liefern. Hierzu zählen u.a. Bloomberg News, Emittentenmitteilungen, Zeitungsmeldungen, Aktienkursinformationen oder Einschätzungen durch Wirtschaftsanalysten. Der Aspekt der gezielten und personalisierten Informationssammlung und Konsolidierung nimmt für den Anleger einen immer wichtiger werdenden Stellenwert ein. Zwar gibt es viele Suchmaschinen, die das Auffinden von Finanzinformationen erleichtern, jedoch liefern Anfragen der Anleger häufig Lösungslisten mit vielen Treffern, deren Qualität sehr unterschiedlich sein kann. Finanzinformationsplattformen wie Bluebull bieten dem Investor die Möglichkeit, verschiedenste Bedürfnisse in einer Portallösung abzubilden. Man findet dort neben aktuellen Marktinformationen auch Analyse- und Berechnungstools, Emissionskalender sowie Top-/Flop- Listen. Die vielfältig online verfügbare Produktauswahl mit einer großen Produktvielfalt (z.B. Aktien, Fonds, Derivate) erschwert es dem Investor dabei allerdings, die für seine individuelle Investitionsentscheidung (z.B. Kauf einer bestimmten DAX-Aktie) benötigten Informationen zusammenzustellen. Auch der Aspekt der Empfehlung oder Einschätzung als Unterstützungskomponente bei einer Finanzentscheidung ist durch diese Art von Plattformen nicht gewährleistet. In Banken kann der Kunde einen Berater um Empfehlungen bitten, aber wo ist im Internet ein Berater, der auf die individuellen Wünsche der Kunden eingehen kann? Es ist notwendig, die Vorstellungen oder Präferenzen (wie z.B. risikoscheuer oder wachstumsorientierter Anleger

mit mittlerem Anlagehorizont) des Investors zu kennen, um ihm personalisiert Investitionsvorschläge unterbreiten zu können.

1.2 Plichtenheft

1.2.1 Zielbestimmung

FIPs ist ein web-basiertes Finanz-Informations-Portal, welches dem Benutzer anhand seiner persönlichen Daten Kauf- und Verkaufsvorschläge von Finanzprodukten unterbreiten soll.
Des Weiteren sollen Nachrichten zu ausgewählten Finanzprodukten zur Verfügung gestellt werden. Diese sollen mit Hilfe von Textmining-Methoden aus diversen Quellen (aus dem Internet) extrahiert werden und in die Bewertung der Finanzprodukte einfließen.

Musskriterien

Die Architektur der Finanz Research Infothek mit Fokussierung auf Expertenfilter und Entscheidungskomponenten soll entworfen werden. Des Weiteren sollen Teile von FIPs implementiert und dokumentiert werden.
FIPs soll in Betrieb genommen, getestet und bewertet werden.

Wunschkriterien

Das gesamte System soll im Internet verfügbar sein. Der Benutzer soll ein persönliches Profil anlegen und verwalten können, auf dessen Grundlage die Kauf- und Verkaufsentscheidungen getroffen werden. Das System speichert die Kundendaten und Informationen zu Finanzprodukten in einer Datenbank.
Nachrichten zu speziellen Finanzprodukten sollen mit Hilfe passender Verarbeitungsmethoden (Wrapper) extrahiert und in die Bewertung der Finanzprodukte einfließen.
Um die zu den Kundenprofilen passenden Finanzprodukte den Kunden zu empfehlen, sollen geeignete Entscheidungsmethoden verwendet werden.
Die Kundenprofile werden anhand eines Fragebogens erstellt und sollen durch die spätere Benutzung an die Kundenwünsche weiter angepasst werden. Das System soll seine Entscheidungen verbessern und adaptive Systemveränderungen durchführen.
Es sollen diverse Finanzprodukte bewertet und zur Verfügung gestellt werden, allerdings wird hierbei zunächst das Augenmerk nur auf DAX-Aktien gelegt. Bei Änderungen der Bewertung von Finanzprodukten anhand von Nachrichten sollen die Kunden, welche diese Finanzprodukte in ihrem Portfolio haben, per Email über die Änderungen benachrichtigt werden.

Abgrenzungskriterien

In den Portfolios der Kunden sollen lediglich die einzelnen Finanzanlagen aufgeführt werden, in die der Kunde investiert, nicht jedoch Anzahl und Vermögen.

1.2.2 Produkteinsatz

Anwendungsbereiche

FIPs ist für Finanzanleger aller Art gedacht, die sich mit Hilfe des Internets über Finanzprodukte informieren und beraten lassen möchten. Es soll die Informationen schnell und einfach zur Verfügung stellen, so dass der geneigte Anleger nicht mehr selbst nach Meldungen suchen und muss. So kann der Benutzer die Entscheidung über Kauf und Verkauf von Finanzprodukten schneller und flexibler gestalten.

Anwendergruppe

Jeder interessierte Anwender kann das System benutzen. Die Anwender müssen grundlegende Kenntnisse im Umgang mit einem Internet-Browser besitzen.

1.2.3 Produktübersicht

Der Systemkern, d.h. die Kernfunktionalität ohne die Verwaltung der Kunden, besteht im Wesentlichen aus den Anwendungsfällen der drei Akteure Admin, KI und Robot (s. Abb. 1.1, Abb. 1.2)
Der Administrator nimmt die wesentlichen Einstellungen am System vor, der Robot ist dafür zuständig aus verschiedenen Quellen zum einen Daten über die Finanzprodukte und zum anderen Finanznachrichten zu holen. Diese Finanznachrichten werden dann von der KI kategorisiert und bewertet.

Akteur Admin:

Das Anwendungsfalldiagramm des Administrators zeigt die wesentlichen Konfigurationsmöglichkeiten des Systemkerns. Der Admin verwaltet die Finanzprodukte und legt damit fest welche Finanzprodukte dem System bekannt sein sollen. Dazu kann er neue Finanzprodukte aufnehmen oder bestehende ändern oder löschen. Finanzprodukte sind hier Aktien, die jeweils einem Unternehmen zugeordnet sind. Eine weitere Aufgabe des Administrators ist die Konfiguration des Robots. Hier werden die verschiedenen Informationsquellen für die Finanzprodukte und Nachrichten festgelegt. Weiter wird das Intervall bestimmt, in dem der Robot die Quellen auf neue Daten überprüfen soll. Schließlich konfiguriert der Admin noch die KI. Das bedeutet, er legt die Kategorien fest, in welche die KI die Nachrichten später einordnet und gibt einen Satz von Entscheidungsregeln vor, welche die KI zur Entscheidungsfindung benutzt.

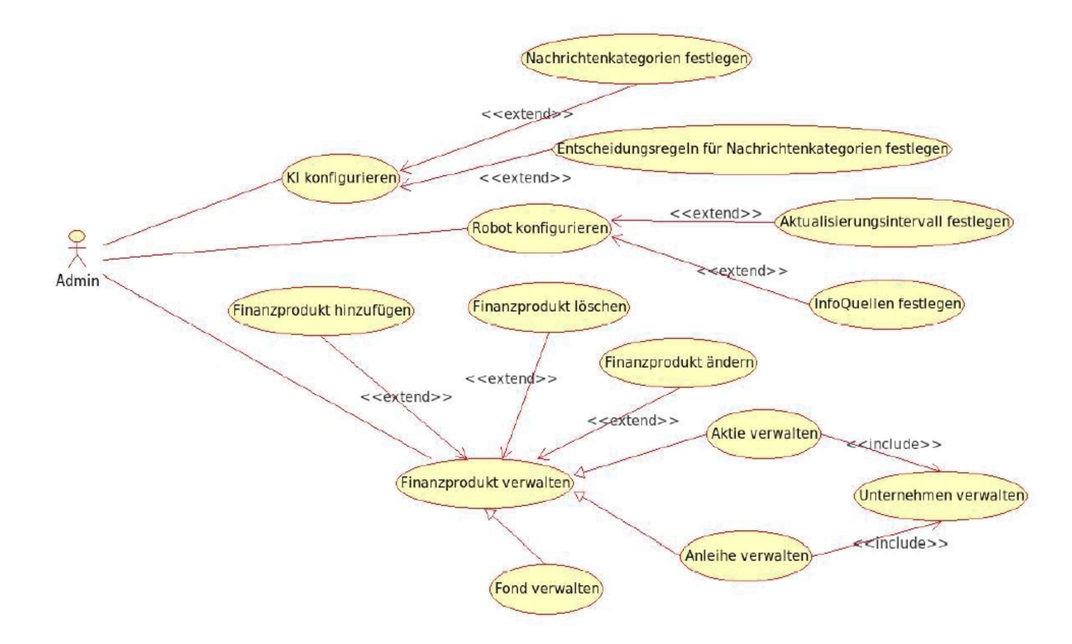


Abbildung 1.1: Admin Anwendungsfalldiagramm

Akteur Robot:

Der Robot holt die Daten eines Finanzproduktes, indem er einen dafür passenden Wrapper startet, bei Aktien also einen Aktienwrapper. Dieser Wrapper liest zuerst seine Konfiguration, extrahiert die entsprechenden Daten aus der Quelle und speichert diese Ergebnisse schließlich in der Datenbank ab. Ähnlich verhält es sich, wenn der Robot Nachrichten holt. Er startet dafür den Nachrichtenwrapper, der entweder Nachrichten aus einem HTML Format oder aus RSS-Feeds extrahiert.

Akteur KI:

Die KI ist dafür zuständig, die vom Robot geholten Nachrichten zu kategorisieren und zu bewerten. Dafür benutzt sie die vom Admin vorgegebenen Entscheidungsregeln. Weiter ist die Aufgabe der KI die Finanzprodukte hinsichtlich Chancen und Risiken zu bewerten, wobei sie auf fundamentale Bewertungskriterien und auf die extrahierten Daten aus den textuellen Nachrichten zurückgreift. Die Bewertung der Finanzprodukte beinhaltet hier die Bewertung der Aktien.

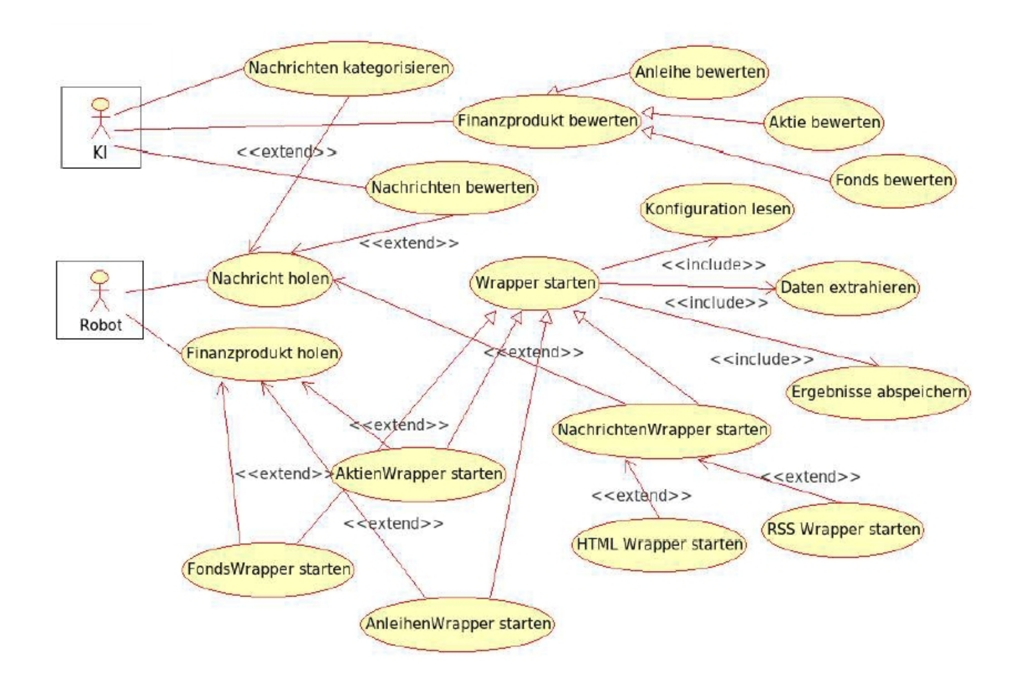


Abbildung 1.2: KI Anwendungsfalldiagramm

1.2.4 Produktfunktionen

Akteur User:

Anwendungsfälle
Neuen User anlegen Durch diese Funktion kann ein neuer User-Account erstellt werden. Wenn der User auf den Link 'Neuer User' klickt, gelangt er in das Fenster, in dem er seine persönlichen Daten (z.B. Vorname, Name) eingeben muss. Zudem muss er auch noch unsere AGB anerkennen. Durch Betätigen des Buttons 'Weiter' öffnet sich ein neues Fenster, in dem er einen Fragebogen zu seiner Einschätzung der eigenen finanziellen Situation ausfüllen soll. Dabei müssen Fragen wie 'Wie hoch ist ihr Nettojahreseinkommen?' beantwortet werden. Mit einem erneuten Klick auf den Button 'Weiter' gelangt der Benutzer zu einem Fenster, in dem er aus einer Liste von Finanzprodukten oder durch Angabe der WKN sein Portfolio zusammenstellen kann. Einen Klick auf den 'Weiter'-Button bestätigt die Eingaben in dem Fenster und führt zur 'Bestätigung ihrer Anmeldung'. Sobald man auf 'OK' geklickt hat, gelangt man in das 'Willkommen'-Fenster, in dem man sich mit dem zugeschickten Zugangsdaten einloggen kann. Ab dem Fenster 'Persönliche Daten' bis zum Fenster 'Portfoliodetails' besteht die Möglichkeit, in das jeweils vorherige Fenster durch Betätigen des Buttons 'Zurück' zu gelangen.

Einloggen Nach Eingabe des Login und des dazu gehörenden Passworts führt das Betätigen des Button 'Login' in das Fenster 'Overview'. Dort wird sein Portfolio mit Links zu den jeweiligen Firmen der Wertpapiere angezeigt. Im Navigationsmenü auf der linken Seite kann man folgende Aktionen ausführen:

1. persönliche Daten bearbeiten
2. Portfolio ändern
3. Fragebogen bearbeiten
4. Links zu Informationsquellen

Durch Ausführen der Aktion 'Empfehlung berechnen' erhält der User gemäß seinen Angaben eine Liste mit Empfehlungen für den Kauf von Wertpapieren.

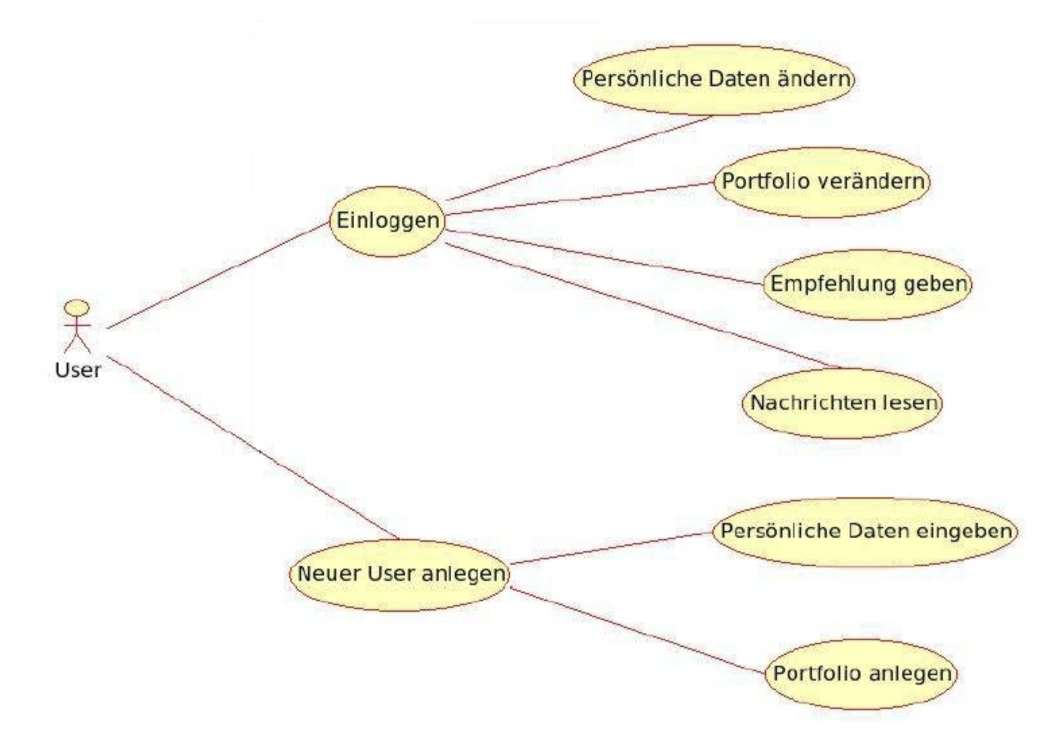


Abbildung 1.3: User Anwendungsfalldiagramm

1.3 Organisatorisches

1.3.1 Gruppentreffen

Die Gruppentreffen finden zwei Mal wöchentlich im Raum 205 in der Otto-Hahn-Straße 16 statt.
Im ersten Semester treffen wir uns am Montag und Donnerstag und im zweiten Semester am Dienstag und Donnerstag jeweils von 8.15 Uhr - 9.45 Uhr.
Die Sitzungsprotokolle befinden sich im Anhang.

1.3.2 Tools

Folgende Tools kommen bei der Umsetzung von FIPs zum Einsatz:

Betriebssysteme:
Debian
Windows 2000/XP

Programmiersprache:
Java 2 Runtime Environment, Second Edition 1.4.2

Servlet/JSP Container:
Apache Tomcat 5.0

Datenbank:
PostgreSQL 7.4.7

Softwarebibliothek:
GATE 3.0

RSS-Reader:
RSS-Owl

1.4 Zeitplan

Hier ist der zeitlich geplante Ablauf des ersten Semesters zu sehen.

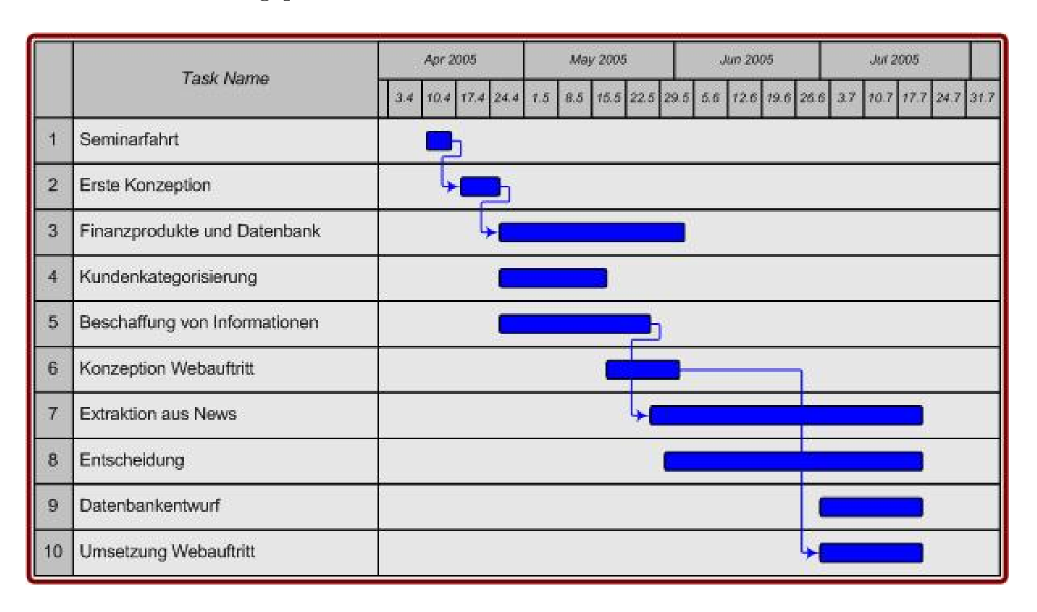


Abbildung 1.4: Zeitplan für das erste Semester

Für das zweite Semester sieht der Zeitplan wie folgt aus.

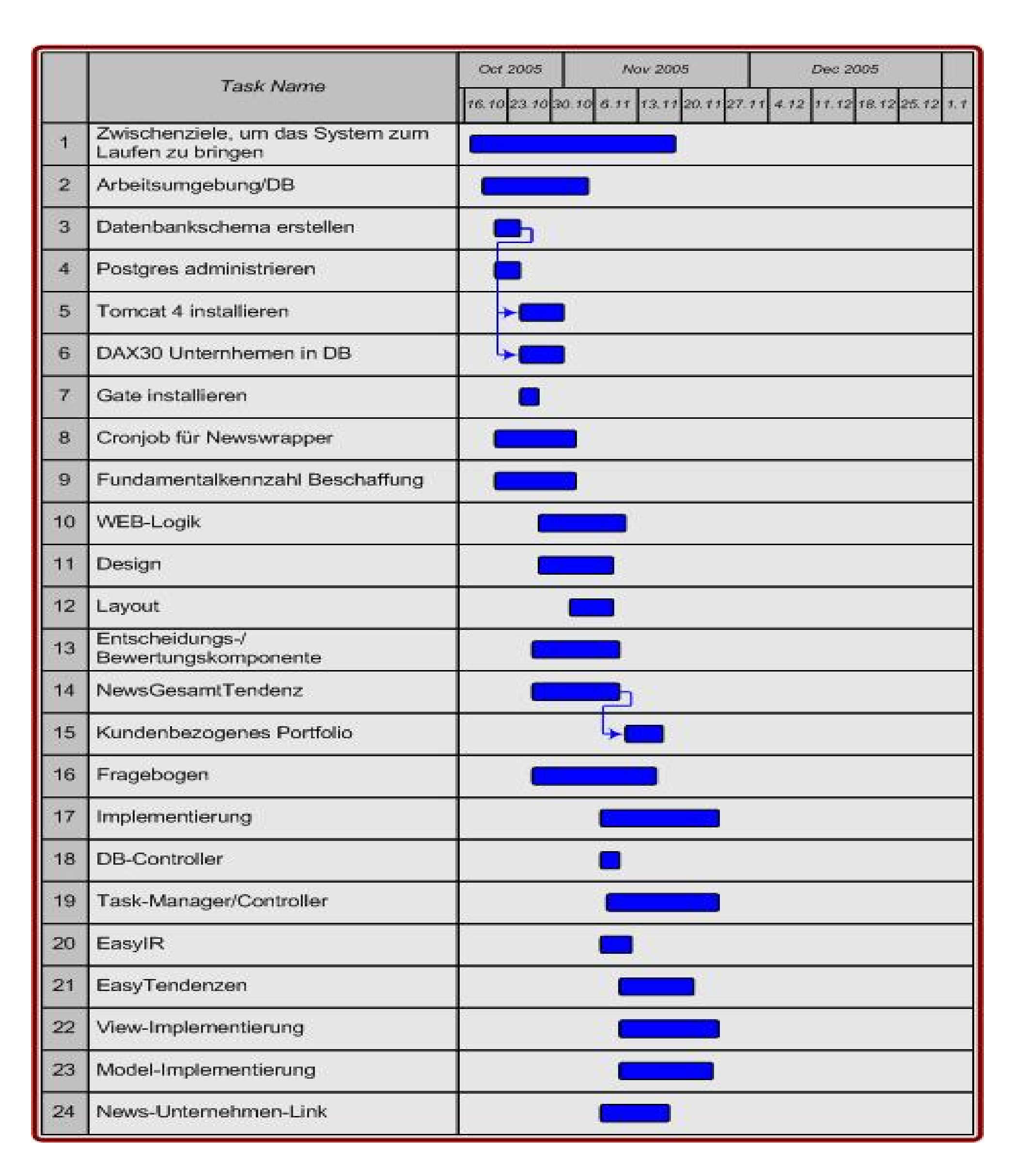


Abbildung 1.5: Zeitplan für das zweite Semester (erste Hälfte)

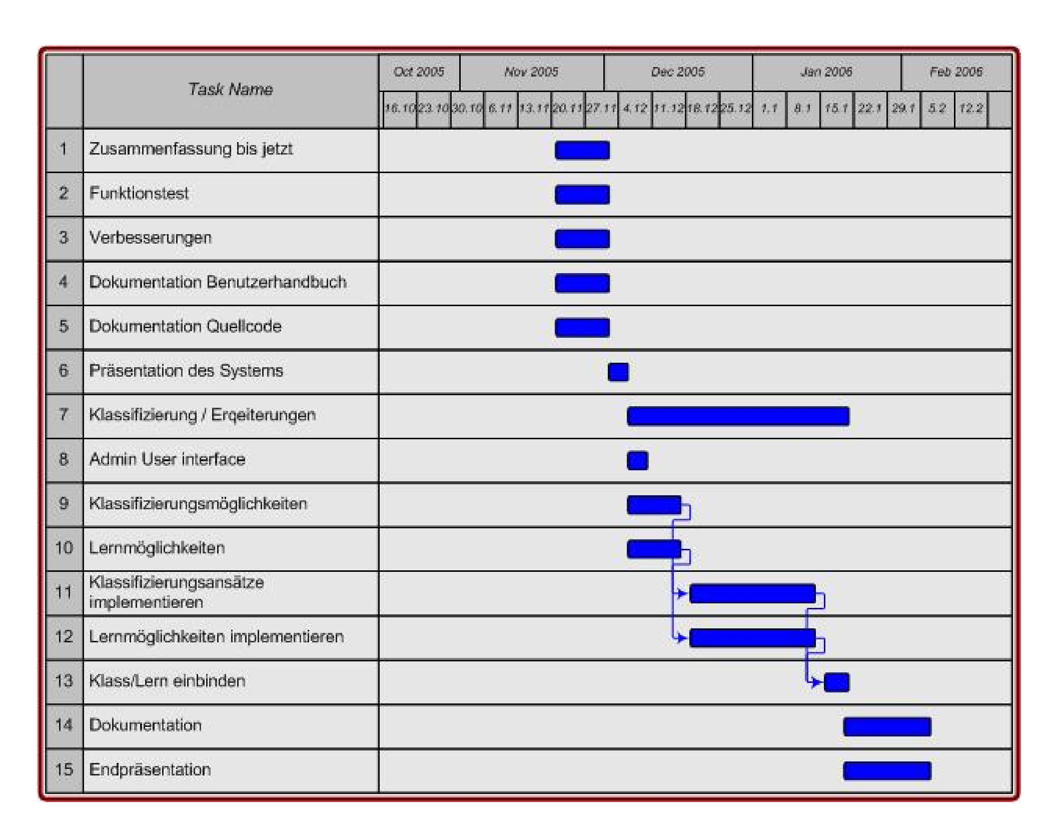


Abbildung 1.6: Zeitplan für das zweite Semester (zweite Hälfte)

Kapitel 2

Erste Schritte

Beim ersten offiziellen Treffen unserer Projektgruppe werden die Themen für die Seminarphase verteilt. Diese Seminarphase findet vom 6. bis 8. April 2005 im 'Haus Nordhelle' in Meinerzhagen statt.
Dort treffen wir auch zum ersten Mal auf unsere Partner der Entory AG, die uns ihr Unternehmen und die geforderten Leistungen an die Projektgruppe vorstellen.
Die Seminarphase dient in erste Linie dazu, sich in die Thematik einzuarbeiten, aber auch, sich untereinander kennenzulernen.
Die Themengebiete umfassen eine Einführung ins Finanz- und Informationsmanagement sowie in Java, J2EE und mit Java verbunden Tools und Softwarebibliotheken. Ein weiterer Bereich umfasst die Grundlagen, Zielsetzungen und Andwendungsgebiete des Text-Minings und bereits vorhandene Werkzeuge für die Problematik. Weitere Vorträge behandeln Graphenbasierte Systeme, Gebiete der Entscheidungstheorie und Präferenzen und Wissen. Der abschließende Vortrag gibt einen Einblick in das Projektmanagement.
Im Extra Anhang 'Seminarausareitungen' befindet sich eine Übersicht über die Seminarvorträge und die Ausarbeitungen der einzelnen Vorträge.

In unseren ersten Projektgruppentreffen nach der vorbereitenden Seminarphase beginnen wir, uns die Ziele unserer PG zu defnieren, einen Zeitplan aufzustellen, was wir bis zum Semesterende und abschliessend zum PG-Ende fertigstellen wollen. Ausßerden wird besprochen, welche Tools eingesetzt werden sollen.

Kapitel 3

Klassifizierung des Kunden

3.1 Einleitung

Um den Kunden Finanz-Empfehlungen zu geben, die auf ihre persönlichen Prioritäten zugeschnitten sind, ist es nötig, die Kunden zu klassifizieren. Dafür müssen Daten erhoben und ausgewertet werden.

3.2 Klassifizierung mittels Kundenvektor

Da man nicht von vorneherein typische Kunden-Kategorien kennt, und auch noch keine Menge typischer Kunden vorgegeben hat, die man zu Clustern zusammenfassen könnte, wird eine möglichst feine Einteilung vorgenommen, die dann mittels Fuzzy-Schnitten den passenden Finanzprodukten zugeordnet werden können.

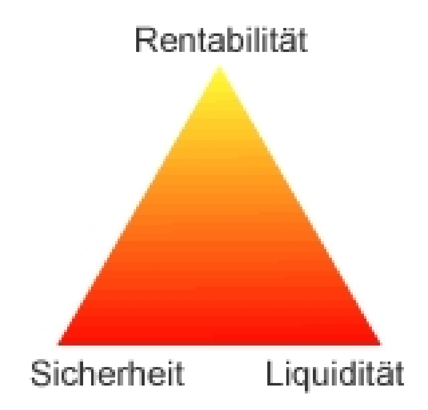


Abbildung 3.1: magisches Dreieck

Das Magische Dreieck der Finanzwelt besagt, dass jedes Finanzprodukt sich

im Spannungsfeld zwischen Verfügbarkeit, Sicherheit und Rendite befindet. Ein Produkt, was möglichst sicher und ständig verfügbar ist, bringt normalerweise keine hohe Rendite; ein Produkt, was relativ sicher und rentabel ist, ist normalerweise nicht kurzfristig verfügbar; usw. Wir versuchen nun, die Prioritäten des Kunden ebenfalls in diesem Dreieck einzuordnen. Jeder Kunde wird gerne eine hohe Rendite haben wollen, aber die Bereitschaft, dafür ein Risiko einzugehen oder sich langfristig festzulegen, ist unterschiedlich ausgeprägt.

Fragenfindung

Mit Hilfe eines Fragebogens versuchen wir nun, das Sicherheitsbedürfnis des Kunden zu ermitteln, um ihn in das Magische Dreieck einordnen zu können. In die Berechnung der Sicherheitspunkte fließen folgende Aspekte ein:

- das Alter des Kunden (0-5 Punkte). Ein Kunde unter 30 kann z.B. vielleicht noch eher ein Risiko eingehen, einem Kunden über 90 ist die Sicherheit der Anlage sicher auch nicht so wichtig wie einem 50jährigen.
- das Nettoeinkommen des Kunden (abzüglich Miete und Versicherungen) (0-20 Punkte). Ein Millionär kann eher ein Risiko bei der Geldanlage eingehen als jemand, der von dem größten Teil seines Geldes alltägliche Ausgaben bestreiten muss.
- die Dauer, für die finanzielle Notfallreserven vorhanden sind (0-20 Punkte). Wer keine Reserven hat, sollte kein Risiko eingehen, wer Reserven für mehrere Jahre hat, kann ruhig spekulieren.
- die gewünschte Anlagesumme (0-20 Punkte). Wer Millionen anlegen will, ist wohl eher bereit, ein Risiko einzugehen.
- maximal hinzunehmender Verlust (0-20 Punkte). Wer bereit ist, auch große Verluste (im Vergleich zur Anlagesumme) hinzunehmen, ist risikofreudiger.
- geplante Anlage zum Teil aus Krediten (0-5 Punkte) ? Wenn ja, sollte man kein großes Risiko eingehen.
- Erfahrungen mit Wertpapieren (0-15 Punkte). Wer noch unerfahren ist, sollte besser sicherere Anlageformen empfohlen bekommen.
- Häufigkeit der Systembenutzung (0-6 Punkte). Ein Kunde, der nur selten auf das System zugreift, kann bei großen Kursschwankungen nicht schnell reagieren und geht besser wenig Risiko ein.
- derzeitiges Portfolio (0-9 Punkte). Daran, welche Finanzprodukte der Kunde schon hat, lässt sich seine Risikofreudigkeit auch einschätzen (wer mit Derivaten handelt, ist risikofreudiger als jemand, der nur ein Sparbuch besitzt).
- eigene Einschätzung der Risikofreudigkeit (0-100 Punkte).

Außerdem werden im Fragebogen personenbezogene Daten zur Registrierung sowie der gewünschte Anlagehorizont abgefragt.

Aussehen des Fragebogens

Auswertung der Angaben

Die gewichtete Summe der Sicherheitspunkte sowie die gewünschte Verfügbarkeit bestimmen nun einen Ort im Magischen Dreieck, für den die mögliche Rendite errechnet wird. Diese drei Werte zusammen ergeben den Kundenvektor (Array aus [Sicherheit][Verfügbarkeit][Rendite]).

Mittels Fuzzy-Logik wird der Kunde anhand dieses Vektors mit bestimmten Zugehörigkeitswerten verschiedenen Klassen zugeordnet, die durch Trapezfunktionen der Werte Sicherheit, Rendite und Verfügbarkeit dargestellt werden können.

Sicherheit

Der Wert für die gewünschte Sicherheit (zwischen 0 und 100) errechnet sich aus den Angaben im Fragebogen (siehe Abbildung 4.2) und wird mit der eigenen Risikoeinschätzung des Kunden abgeglichen.

Verfügbarkeit

Der Wert für die gewünschte Verfügbarkeit (zwischen 0 und 100) wird dem Fragebogen entnommen (Anlagehorizont).

Rendite

Der Wert für die Rendite (zwischen 0 und 100) ergibt sich durch das magische Dreieck automatisch aus denen für Sicherheit und Verfügbarkeit.

Fragebogen zur Kundenklassifizierung
sowie
Verteilung von Sicherheitspunkten

Personenbezogene Daten	**Sicherheitspunkte = ⌈Summe aller Punkte hier : 1,2⌉**
* Anrede	
* Name	
* Vorname	
* E-Mail	
* E-Mail (Wiederholung)	
* Geburtsjahr	*Alter = (aktuelles Jahr – Geburtsjahr) <30 oder >90:* **0**
	Alter zwischen 30-60: **(Alter – 30) : 6**
	Alter zwischen 60-90: **10 - ((Alter – 30) : 6)**
Haushalts-Nettojahreseinkommen abzüglich Miete + Versicherungen (in €)	**723/Wurzel(Einkommen) – Einkommen/6000 + 14,43** (*vgl. Graph unten)
______________€	
Wie viele Personen leben in Ihrem Haushalt ?	*pro zusätzl. Erwachsenem (Kind) werden vom*
___ Erwachsene ___ Kinder	*Einkommen oben 4000 (3000) € abgezogen*
Notfallreserven für	
o unter 2 Monate	**andere Empfehlung**
o 2-3 Monate	**20**
o 3-6 Monate	**10**
o >6 Monate	**0**
Anlagesumme	*Anlagesumme < 2500 €:* **andere Empfehlung**
______________€	*Anlagesumme > 100000 €:* **0** *sonst:* **50000 : Anlagesumme**
Welchen Verlust sind Sie bereit, maximal hinzunehmen ?	*(Anlagesumme : maximaler Verlust) < 2:* **0**
______________€	*sonst:* **(Anlagesumme : maximaler Verlust – 2) : 5**
Stammen die Mittel der Anlagesumme aus Krediten?	
o ja	**5**
o nein	**0**
o teilweise	**3**

Abbildung 3.2: Fragebogen mit zugeordneten Sicherheitspunkten (obere Hälfte)

Anlagehorizont	
o kurzfristig	
o halbes bis ein Jahr	
o 1-2 Jahre	
o 2-5 Jahre	
o 5-10 Jahre	
o >10 Jahre	
Erfahrungen mit Wertpapieren	
o keine	15
o bis 2 Jahre	11
o 2-5 Jahre	7
o 5-10 Jahre	3
o >10 Jahre	0
Wie häufig wollen Sie unser System durchschnittlich benutzen?	
o täglich	0
o mehrmals wöchentlich	2
o mehrmals monatlich	4
o seltener	6
Wie schätzen Sie Ihre Risikofreudigkeit für die geplante Anlage ein?	
____________________||____________________	
kein Risiko sehr hohes Risiko	
In welche Anlageformen investieren Sie derzeit schon?	**Mittel aus ...**
(Mehrfachnennung möglich)	
[] keine	9
[] Aktien	1
[] Anleihen	5
[] Immobilien	4
[] Devisen	1
[] Sparbuch	9
[] Fonds	3
[] Rohstoffe	1
[] Derivate	0
[] andere	0

Graph zum Einkommen

20.0 18.0 16.0 14.0 12.0 10.0 8.0 6.0 4.0 2.0 0.0

10000 20000 30000 40000 50000 60000 70000 80000 90000 100000

Abbildung 3.3: Fragebogen mit zugeordneten Sicherheitspunkten (untere Hälfte)

Kapitel 4

Klassifizierung von Finanzprodukten (techn. Analyse)

4.1 Einleitung

Bei der Klassifizierung von Finanzprodukten betrachten wir aufgrund der grossen Menge unterschiedlicher Produkte nur die Aktie. Desweiteren beschränken wir uns auf die Aktien, die im DAX30 vertreten sind. In diesem Kapitel wird nun beschrieben, wie diese dreißig Aktien in verschiedene Klassen eingeordnet werden, um eine kompakte Repräsentation zu erhalten. Basierend auf diesen Daten sollen dem Kunden Kaufempfehlungen gegeben werden.

Die zu einer Aktie gehörenden Klassen sind "Performer" und "Sicherheit". In der Klasse "Performer" sind die Ausprägungen "Outperformer", "Marketperformer" und "Underperformer" enthalten. Sie gibt an, wie eine Aktie am Markt positioniert ist. Die Ausprägung "Outperformer" beispielsweise deutet auf eine gute Position der Aktie am Markt hin und verspricht somit eine positive Entwicklung bzgl. des Aktienkurses. Dementsprechend sind "Marketperformer" diejenigen Unternehmen, deren Kurse sich weder positiv noch negativ entwickeln. "Underperformer" weisen schlechtere Zahlen als der Branchen- oder Marktdurchschnitt vor.

Die Klasse "Sicherheit" gibt an, wie hoch das Risiko eines (Geld-)Verlustes bzgl. der betrachteten Aktie ist. Sie ist unterteilt in die vier Klassen "mittel", "gering / mittel","gering" und "sehr gering". Aufgrund des spekulativen Charakters von Aktien sind nur solche Werte bzgl. ihrer Sicherheit möglich. Die Einschätzung "sehr gering" deutet dabei auf einen stark schwankenden Kurs hin, bei dem eventuell ein hoher Verlust entstehen könnte. Die beste Einschätzung ist bei einer Aktie "mittel" und garantiert eine geringe Wahrscheinlichkeit eines Verlustes.

Anders als bei der Klassifizierung des Kunden, bei der drei Klassen eine Rolle spielen, werden bei der Aktie nur zwei Klassen berechnet. Da nur ein Finanzprodukt behandelt wird, ist die Frage der Verfügbarkeit irrelevant, da sie für alle Aktien gleich ist.

4.2 Fundamentale Kennzahlen

Im Folgenden werden die 10 unternehmensbezogenen Kennzahlen zur fundamentalen Bewertung von Aktien beschrieben. Sie lauten: KGV, PEG, DIV, KCV, KBV, MU, CM, EBIT, EBITDA und EKR.

KGV Das Kurs-Gewinn-Verhältnis stellt den Gewinn eines Unternehmens mit der aktuellen Börsenbewertung in Verhältnis. Diese Rentabilitätskennziffer ist eines der gängigsten Instrumente zur Beurteilung von Aktien. Dazu wird das KGV einer Aktie mit dem Durchschnitts-KGV des gesamten Marktes oder Branche verglichen. Ist das KGV der Aktie geringer als das Durchschnitts-KGV, so deutet dies auf eine preiswerte Aktie hin. Allgemein kann man sagen, dass ein niedriger das KGV auf eine günstige Bewertung der Aktie hinweist.

PEG Die Kennzahl "Price-Earning to Growth-Ratio" setzt das KGV eines Geschäftsjahres in Relation zum erwarteten Gewinnwachstum im kommenden Geschäftsjahr. Diese Kennzahl wird vor allem bei Wachstumswerten zur Bewertung eingesetzt, insbesondere für Unternehmen, die wertsteigernde Wachstumschancen besitzen. Wachstum hat einen positiven Einfluß auf den Unternehmenswert.

- PEG-Ratio < 1: Unterbewertung, da KGV geringer als die Wachstumsrate
- PEG-Ratio > 1: Überbewertung
- PEG-Ratio $= 1$: faire Bewertung

Allerdings muss das PEG eines Unternehmens sinnvollerweise mit dem Branchendurchschnitt verglichen werden. Wenn eine bestimmte Aktie eine PEG-Ratio von 1,1 aufweist und die PEG-Ratio des Branchendurchschnitts 1,4 beträgt, dann kann man nicht in isolierter Betrachtung des Unternehmens von einer Überbewertung ausgehen.

DIV Die Dividendenrendite (in %) ist eine Kennzahl zur Bewertung und zum direkten Vergleich von Aktien. Die Kennzahl setzt die vom Unternehmen gezahlte Dividende mit dem Kurs der Aktien ins Verhältnis. Dabei können die Berechnungen sowohl auf der Basis der momentan gezahlten Dividende, als auch auf Basis von erwarteten künftigen Dividenden erfolgen. Allgemein kann man sagen, dass je höher die Dividendenrendite ist, umso höher ist auch die Dividende, die der Anleger bekommt. Allerdings kann das Kapital der Gewinnausschüttung vom Unternehmen nicht für wachstumssteigernde Maßnahmen eingesetzt werden.

KCV Der Kurs-Cash-Flow ist der Quotient aus dem Aktienkurs und Cash-Flow je Aktie (Aktienkurs/Cashflow). Der Cashflow (auch Umsatzüberschuß, Finanzüberschuß) ist der Nettozugang an liquiden Mitteln aus der Umsatztätigkeit und sonstigen laufenden Aktivitäten während einer Periode. Der Cashflow an sich ist schon eine sehr gängige und aussagestarke Kennzahl.

Nach der Definition des Cashflow ist das KCV nun eine liquiditätsorientierte Kennzahl. Sie wird zur Bewertung der Entwicklung der Ertragskraft einer Unternehmung in der Zukunft sowie zum Vergleich verschiedener Unternehmungen miteinander verwendet. Falls das KGV einer Aktiengesellschaft aufgrund von Verlusten der Aktiengesellschaft nicht errechnet werden kann, dann wird das KCV angewendet. Je niedriger das Verhältnis ist, desto günstiger ist die Aktie bewertet.

KBV Der Buchwert einer Aktie entspricht dem Eigenkapital dividiert durch die Anzahl der Aktien und bezeichnet den bilanziellen Wert des Unternehmensteils, der dem Anleger in Form einer Aktie gehört. Das Kurs-Buchwert-Verhältnis wird zur Beurteilung der Substanz eines Unternehmens verwendet. Dazu wird der Kurs einer einzelnen Aktie in Relation zu ihrem Buchwert gestellt. Je niedriger das KBV, desto preiswürdiger ist die Aktie.

MU Die Kennzahl "Marktkapitalisierung pro Umsatz" setzt die Marktkapitalisierung ins Verhältnis zum Umsatz des Unternehmens im Geschäftsjahr. Bei der Berechnung wird die Marktkapitalisierung des vergangenen Geschäftsjahres ins Verhältnis zum Umsatz des letzten Geschäftsjahres gesetzt. Sie sagt aus, wie hoch ein Euro Umsatz an der Börse bewertet wird. Je höher diese Kennzahl ist, desto höher wird das Unternehmen an der Börse bewertet. Zum Beispiel sagt ein Wert von 0,50 aus, daß ein Euro Umsatz zur Zeit mit 50 Cents an der Börse bewertet wird. Je niedriger diese Kennzahl ist, desto günstiger ist die Aktie bewertet.

CM Die Cash-Flow-Marge ist eine Kennzahl für die operative Unternehmensrentabilität. Sie gibt an, wie viel Prozent der Umsatzerlöse dem Unternehmen zur Investitionsfinanzierung, Schuldentilgung und Dividendenzahlung frei zur Verfügung stehen. Sie ist Maßstab für die Ertrags- und Selbstfinanzierungskraft des Unternehmens. Je höher die Kennzahl ist, desto höher ist die Unternehmensrentabilität.

EBIT Die EBIT-Marge ist eine operative Unternehmenskennzahl. Sie berechnet sich aus der Relation des EBIT zum Umsatz. Sie ist als relative Kennzahl prädestiniert um die EBIT-Ertragskraft verschiedener Gesellschaften miteinander zu vergleichen.

EBIT (Earnings before interest and taxes) wird aus dem Jahresüberschuß vor Steuern, Zinsergebnis und vor außerordentlichem Ergebnis berechnet. Durch die Eliminierung dieser genannten Faktoren, erhält man eine vergleichbarere Aussage über die eigentliche operative Ertragskraft einer Unternehmung und zwar

unabhängig von der individuellen Kapitalstruktur. Bei Verwendung des Jahresüberschußes bzw. der Netto-Umsatzrendite schneiden nämlich Unternehmen mit einer höheren Eigenkapitalquote aufgrund geringerer Fremdkapitalkosten tendenziell besser ab.

EBITDA Die EBITDA-Marge (in %) ist eine operative Unternehmenskennzahl. Sie berechnet sich aus der Relation des EBITDA zum Umsatz. Sie ist als relative Kennzahl prädestiniert, um die EBITDA-Ertragskraft verschiedener Gesellschaften miteinander zu vergleichen.

EBITDA (earnings before interests, taxes, depreciation and amortization) setzt sich aus dem Jahresüberschuss vor Steuern, dem Zinsergebnis und den Abschreibungen des Unternehmens zusammen. Das EBITDA ist eine international weitverbreitete und eine der aussagekräftigsten Erfolgskennzahlen, um die operative Ertragskraft einer Gesellschaft zu beurteilen. Da international betrachtet die Gesellschaften unter unterschiedlichen Gesetzgebungen bilanzieren, ermöglicht die Kennzahl EBITDA aufbauend auf dem EBIT aussagekräftigere Vergleiche der operativen Ertragskraft als man durch den ausgewiesenen Jahresüberschuß erhält. Beispielsweise weisen investitionsfreudige Unternehmen hohe ergebnismindernde Abschreibungen und damit einen geringeren Jahresüberschuß als weniger investitionsfreudige Unternehmen auf. Somit hat das EBITDA einen gewissen Bereinigungscharakter.

EKR Die Eigenkapitalrendite (in %) entspricht der Kapitalrentabilität eines Unternehmens. Sie errechnet sich aus dem Jahresüberschuß dividiert durch das eingesetzte Eigenkapital. Sie gibt die Verzinsung des Eigenkapitals an und ist deswegen vor allem aus Sicht der Aktionäre wichtig. Im Vergleich zu anderen Unternehmen einer Branche gilt grundsätzlich: Je höher die Eigenkapitalrendite desto positiver fällt eine Bewertung für das Unternehmen aus. Allerdings muß eine relativ geringe Eigenkapitalrendite für sich nicht unbedingt negativ interpretiert werden, falls z.B. die Gesellschaft diese in den letzten Geschäftsjahren sukzessive erhöhen konnte, der Trend also positiv ist. Dann läßt sich hieraus interpretieren, daß das Managment die Ertragssituation in den Griff bekommt.

4.3 Bestimmung des Performers von Finanzprodukten

Der von uns berechnete Performer für eine Aktie wird auf der Basis von Branchendurchschnittswerten (und Standardabweichungen), Kennzahlwerte der Aktie und Nachrichtenbewertungen ermittelt.

Die Branchendurchschnittswerte für jede Kennzahl sind manuell für jede im DAX30 vorkommende Branche bestimmt worden und in der Datenbank gespeichert. Die Werte für jede Kennzahl einer Aktie im DAX30 wurden per HTML-Wrapper in die DB geschrieben. In einem zweistufigen Verfahren wird auf der

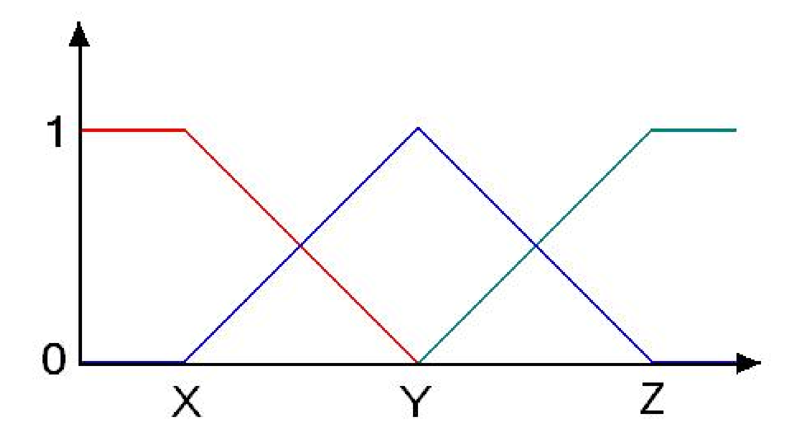


Abbildung 4.1: Fuzzymengen für eine Kennzahl. Rot: $f_N^i(x)$, blau: $f_M^i(x)$ und grün: $f_H^i(x)$

ersten Stufe zunächst ein Basisperformer errechnet. Diesem liegen nur die Fundamentalkennzahlen zugrunde. Da diese Kennzahlen nur jährlich aktualisiert werden, ist eine feinkörnige Prognose erstmal nicht möglich. Daher wird auf der zweiten Stufe des Verfahrens dieser Basisperformer mit den aktuellen Nachrichtenbewertungen kombiniert. Somit wird z.B. ein Unternehmen, das fundamental gesehen outperformt und innerhalb des Jahres schlechte Schlagzeilen macht, eine Klassifizierung "marketperformer" oder sogar nur "underperformer" erhalten.

Erste Stufe Zunächst wird für jede Kennzahl der betrachteten Aktie festgestellt, ob sie über, unter oder nah an dem Durchschnittswert liegt. Dafür erstellen wir Fuzzymengen

$$f_M^i(x) = \begin{cases} 0 & \text{falls } x < X^i \\ \frac{1}{\sigma * \alpha}(x - X^i) & \text{falls } X^i < x \leq Y^i \\ 1 - \frac{1}{\sigma * \alpha}(x - Y^i) & \text{falls } Y^i < x \leq Z^i \\ 0 & \text{sonst} \end{cases}$$

$$f_N^i(x) = \begin{cases} 1 & \text{falls } x < X^i \\ 1 - \frac{1}{\sigma * \alpha}(x - X^i) & \text{falls } X^i < x \leq Y^i \\ 0 & \text{sonst} \end{cases}$$

$$f_H^i(x) = \begin{cases} 0 & \text{falls } x < Y^i \\ \frac{1}{\sigma * \alpha}(x - Y^i) & \text{falls } Y^i < x \leq Z^i \\ 1 & \text{sonst} \end{cases}$$

mit $i \in K := \{$KGV,PEG,DIV,KCV,KBV,MU,CM,EBIT,EBITDA,EKR$\}$.

Die Mengen $f_M^i(x)$ entsprechen dem Durchschnittswert. Die Mengen $f_N^i(x)$ bzw. $f_H^i(x)$ stehen jeweils für niedrigere bzw. höhere Werte als der Durchschnitt.

In Abbildung 4.1 wird dies für eine Kennzahl i dargestellt. Der Wert der Kennzahl der betrachteten Aktie kann dann eingesetzt werden, um die Zugehörigkeit zu hoch, mittel und niedrig zu erhalten.

Abhängig von der Kennzahl ist nun ein überdurchschnittlicher Wert ein Hinweis auf einen Outperformer oder Underperformer. Beispielsweise ist für die Kennzahl KGV ein überdurchschnittlicher Wert ein Indiz für einen Underperformer und ein unterdurchschnittlicher Wert spricht für einen Outperformer.

Die einzelnen Regeln können der Tabelle in Abbildung 4.2 entnommen werden. Durch die Zugehörigkeitswerte und die Regeln erhalten wir $3 * 10$ Performancewerte (von jeder Kennzahl 3 Performer). Um nun aus den einzelnen Kennzahlen eine Gesamtperformance für diese Aktie zu erhalten, müssen die einzelnen Performanzen zusammengeführt werden. Jetzt ist es allerdings so, dass die einzelnen Kennzahlen nicht uniform mit in die Gesamtperformance einfliessen. Zum Beispiel bestimmt das KGV mehr die Performance als KBV oder MU. Ebenso können Ausreißerperformer (Under- und Outperformer) interessanter sein als Marketperformer, wodurch letztere nicht so stark gewichtet werden können. Die Gewichtung einer Regel wird dann noch unterschieden nach der Sicherheit der Aktie (mittel, gering/mittel, gering, sehr gering; siehe auch Abb.4.2).

Zur Normierung der unterschiedlichen Gewichtungen (Normierungsfaktoren λ_O,λ_M,λ_N), wird für jede Performance (da für jede Performance ja auch Regeln vorhanden sind) das mittlere Gewicht bestimmt. In der Tabelle wird also für den Outperformer das Mittel über die Gewichte von "Regel 3","Regel 6","Regel 7","Regel 12","Regel 15","Regel 18","Regel 19", "Regel 22","Regel 25" und "Regel 28" gebildet.

Für die Gesamtperformance wird die Zugehörigkeit einer Aktie wie folgt bestimmt

$$Zugehörigkeit(q, s) = \frac{1}{N * \lambda_q} \sum_{i \in W} p_q^i * g_{q,s}^i$$

mit $s \in$ {mittel,gering/mittel,gering,sehr gering}
$q \in$ {Outperformer,Marketperformer,Underperformer}
$N \hat{=}$Anzahl Regeln
$p_q^i \hat{=}$Zugehörigkeit zu q bzgl. Kennzahl i
$g_{q,s}^i \hat{=}$Gewicht zu q bei Sicherheitsklasse s bzgl. Kennzahl i
$W \hat{=}$Menge der Regeln

4.3.1 Differenzierung des Performers

Um einer Aktie nun einen konkreten Performer zuzuweisen, werden die Zugehörigkeitswerte betrachtet. Mit Hilfe einiger Testdaten fanden wir heraus, dass bei einigen Aktien die Zugehörigkeitswerte von Under- und Outperformer sehr nah zusammenliegen. Eine eindeutige Zuweisung nach Under- oder Outperformer ist demnach nur schlecht möglich (falls diese Werte größer sind als der Zugehörigkeitswert zu Marketperformer).

Regel			**Gewichtung**
Regel 1	Wenn KGV=Hoch	dann PKGV=Under	0,9
Regel 2	Wenn KGV=Mittel	dann PKGV=Market	0,5
Regel 3	Wenn KGV=Niedrig	dann PKGV=Out	0,9
Regel 4	Wenn PEG=Hoch	dann PPEG=Under	0,7
Regel 5	Wenn PEG=Mittel	dann PPEG=Market	0,6
Regel 6	Wenn PEG=Niedrig	dann PPEG=Out	0,7
Regel 7	Wenn DIV=Hoch	dann PDIV=Out	0,8
Regel 8	Wenn DIV=Mittel	dann PDIV=Market	0,6
Regel 9	Wenn DIV=Niedrig	dann PDIV=Under	0,3
Regel 10	Wenn KCV=Hoch	dann PKCV=Under	0,71
Regel 11	Wenn KCV=Mittel	dann PKCV=Market	0,36
Regel 12	Wenn KCV=Niedrig	dann PKCV=Out	0,71
Regel 13	Wenn KBV=Hoch	dann PKBV=Under	0,34
Regel 14	Wenn KBV=Mittel	dann PKBV=Market	0,3
Regel 15	Wenn KBV=Niedrig	dann PKBV=Out	0,34
Regel 16	Wenn MU=Hoch	dann PMU=Under	0,4
Regel 17	Wenn MU=Mittel	dann PMU=Market	0,32
Regel 18	Wenn MU=Niedrig	dann PMU=Out	0,4
Regel 19	Wenn CM=Hoch	dann PCM=Out	0,8
Regel 20	Wenn CM=Mittel	dann PCM=Market	0,63
Regel 21	Wenn CM=Niedrig	dann PCM=Under	0,8
Regel 22	Wenn EBIT=Hoch	dann PEBIT=Out	0,5
Regel 23	Wenn EBIT=Mittel	dann PEBIT=Market	0,5
Regel 24	Wenn EBIT=Niedrig	dann PEBIT=Under	0,5
Regel 25	Wenn EBITDA=Hoch	dann PEBITDA=Out	0,5
Regel 26	Wenn EBITDA=Mittel	dann PEBITDA=Market	0,5
Regel 27	Wenn EBITDA=Niedrig	dann PEBITDA=Under	0,5
Regel 28	Wenn EKR=Hoch	dann PEKR=Out	0,43
Regel 29	Wenn EKR=Mittel	dann PEKR=Market	0,3
Regel 30	Wenn EKR=Niedrig	dann PEKR=Under	0,43

Abbildung 4.2: Regelmenge für Zuweisung {hoch, mittel,niedrig} $\rightarrow$ {out, market, under} mit Gewichtungen für die Sicherheitsklasse "mittel"

Daher wird der Aktie der Performer Marketperformer zugewiesen, falls die Differenz zwischen den Zugehörigkeitswerten von Out- und Underperformer kleiner als 0,2 ist. Ansonsten wird die Klasse mit dem höchsten Zugehörigkeitswert gewählt.

Zweite Stufe Der in der ersten Stufe berechnete Basisperformer wird nun mit der Gesamttendenz der relevanten Nachrichten kombiniert und es ergibt sich der endgültige Performer, der für die Bewertungs-/Rankingfunktion benötigt wird.

Da die automatische Bewertung von Nachrichten durch das System nicht zufriedenstellend arbeitet (siehe Kapitel 5) greifen wir auf die Nachrichtenbewertungen der Nutzer zurück. Jede Nachricht, die in der Datenbank gespeichert wurde, kann ein Benutzer lesen und anschliessend hinsichtlich der Relevanz und Tendenz für eine Aktie bewerten. Sollte ein Benutzer eine Nachricht für besonders relevant für eine Aktie erachten, so kann er auf einer Skala von 0 bis 100 einen sehr hohen Wert auswählen. Nicht relevante Nachrichten werden mit einer 0 bewertet. Die Bewertung der Tendenz erfolgt ebenfalls auf einer Skala von 0 bis 100. Sollte eine Nachricht eine höchst positive Auswirkung auf eine Aktie haben, so wird ein hoher Wert gewählt. Eine negative Einschätzung ist mit einer niedrigen Zahl verbunden.

Neben den Tendenzen und Relevanzen spielt natürlich auch die Glaubwürdigkeit des jeweiligen Nutzers eine Rolle. Ein Bonussystem soll den Bewertungen eines Nutzers eine Gewichtung geben. Wenn sich die Vorhersagen eines Nutzers bestätigen, dann erhalten seine zukünftigen Bewertungen ein höheres Gewicht und umgekehrt (siehe Kapitel 8).

Die Gesamttendenz ergibt sich aus der normierten Aggregation dieser 3 Faktoren (Tendenz, Relevanz und Kundenstatus) für alle Nachrichten, die für die betrachtete Aktie relevant sind. Diese Gesamttendenz kann dann als Newsperformer interpretiert werden.

Der Gesamtperformer ergibt sich also aus der Kombination von Basisperformer und Newsperformer, wobei die Anzahl der Bewertungen für die relevanten Nachrichten ausschlaggebend für die Gewichtung des Newsperformers ist (maximales Gewicht für den Newsperformer beträgt 75%).

4.4 Berechnung der Sicherheit eines Finanzproduktes

Die Berechnung der Sicherheit eines Finanzproduktes ergibt sich aus den vier Kennzahlen DIV, KBV, EKR und PEG. Diese Werte werden geeignet kombiniert, um einen Sicherheitswert $s \in [0, 2]$ zu erhalten. Die Klassen "mittel", "mittel/gering", "gering" und "sehr gering" sind durch Fuzzymengen auf einem Bereich von 0 bis 2 definiert (siehe Abbildung 4.3).

Die Sicherheitsklasse der Aktie entspricht nun der Klasse, für die die jeweilige Fuzzyfunktion den größten Zugehörigkeitswert bzgl. des Sicherheitswertes

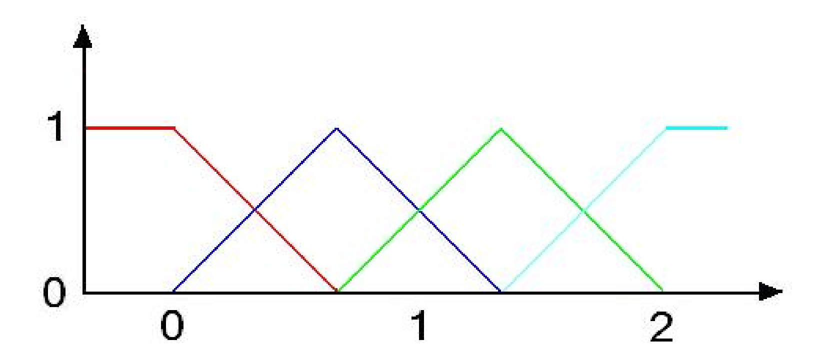


Abbildung 4.3: Fuzzymengen für die Sicherheitsklassen; rot: "sehr gering", dunkelblau: "gering", grün: "mittel/gering", hellblau:"mittel"

besitzt, also

$$Sicherheitsklasse_{Aktie} = argmax_{i \in SK} f_i(s)$$

mit
$SK \hat{=}$ {mittel, mittel/gering, gering, sehr gering}
$f_i \hat{=}$ Fuzzymenge für Sicherheitsklasse i

Bestimmung des Sicherheitswertes Für jede Kennzahl erstellen wir 3 Fuzzymengen (positiv, negativ und neutral), die angeben, inwiefern sich diese Kennzahl auf die Sicherheit auswirkt. Für die Kennzahl DIV ist ein Wert über 2 ein Hinweis auf hohe Sicherheit und ein Wert unter 1 spricht für eine geringe Sicherheit. Die einzelnen Fuzzymengen sind in Abbildung 4.4 definiert.

Nun wird für jede Kennzahl die Klasse mit dem größten Zugehörigkeitswert bestimmt. Den Klassen sind bestimmte Werte zugeordnet. So hat die Klasse "positiv" einen Wert von 2, "neutral" 1 und "negativ" 0. Über diese vier Kennzahlen bestimmen wir nun die durchschnittliche Klasse bzgl. dieser Werte. Der Wert liegt also zwischen 0 und 2. Der Sicherheitswert ergibt sich nun durch die Multiplikation dieses Wertes mit dem Mittel der vier maximalen Zugehörigkeitswerte.

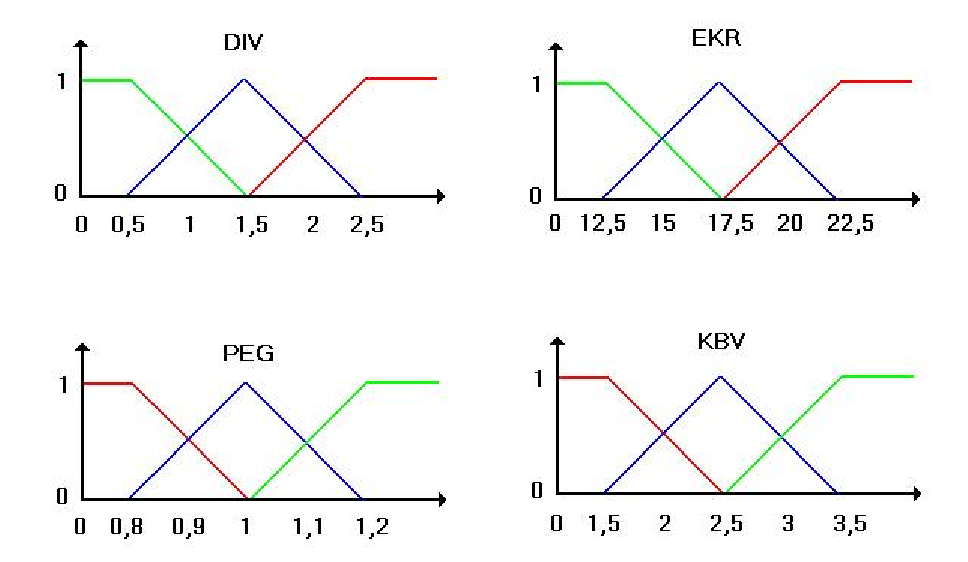


Abbildung 4.4: Fuzzymengen für Sicherheitsbetrachtung; rot: positiv, blau: neutral, grün: negativ

Kapitel 5

Bewertungsansätze für Finanznachrichten

5.1 Überblick

Dieses Kapitel geht einerseits auf die erstellten Bewertungskategorien, denen die Finanznachrichten zugeordnet werden, andererseits auf die angewendeten Verfahren zur Einordnung in diese Kategorien und den dabei aufgetretenen Schwierigkeiten ein.

Um Finanznachrichten überhaupt bewerten zu können, muss zunächst festgestellt werden, was in diesen Nachrichten steht. Dazu werden in Kapitel 5.4 zunächst Methoden beschrieben, die das Preprocessing angehen. Diese sind unabhängig von den späteren Methoden, die in Kapitel 5.5 beschrieben sind.

Kapitel 5.2 zeigt zunächst das Ziel auf, nämlich eine Kategorisierung der Unternehmungen aufgrund der bearbeiteten Nachrichten.
Kapitel 5.3 beschäftigt sich mit der Einbindung der Methoden ins Gesamtsystem und den dabei entstandenen Schwierigkeiten.

5.2 Finanznachrichten in Klassen under-, market- und outperformer

Wie im Kapitel 4.3 beschrieben, werden die Aktien aufgrund ihrer Kennzahlen in drei Klassen eingeteilt: Under-, market-, oder outperformer.

Hier beschäftigen wir uns allerdings nicht mit der Einteilung der Aktien durch ihre Kennzahlen, sondern mit dem Problem, ob wir mit Hilfe von bekannten Methoden herausfinden können,

- dass eine Kongruenz zwischen Nachrichten und der Entwicklung der Aktie besteht

und falls es eine gibt, wollen wir

- die Stimmungen, die bestimmte Nachrichten wiedergeben, in die Klassifikation der Aktie mit einbringen.

Diese Ziele bedingen eine Einteilung der Nachrichten selbst. Nur wenn wir sagen könnten, dass sich eine Nachricht positiv, eine andere negativ oder möglicherweise auch neutral auswirkt, können Rückschlüsse auf Aktienentwicklungen am Markt gezogen werden.

Das Stichwort "Einteilung von Texten in Kategorien" evoziert sofort das Stichwort "Klassifikationsmethoden des Text-Mining" bzw. "Klassifikationsmethoden" überhaupt. Beim Arbeiten mit den Texten lautete unsere Arbeitshypothese:

- Eine Ähnlichkeit der Anzahl respektive Häufigkeit der vorkommenden Wörter impliziert eine Ähnlichkeit der Semantik.

Diese Hypothese erklärt die gewählten Methoden, die eine breite Palette an bekannten Klassifizierungsarten abdecken. Sie reicht von einer einfachen Clusteranalyse á la k-means, bis hinüber zu über- und unüberwachten Lernmethoden aus dem Bereich des maschinellen Lernens, wie zum Beispiel support-vector-machines oder ART 2a-Netze.

Auf jede dieser Methoden wird unten noch einmal genau eingegangen.

5.3 Allgemeines Verfahren, Schwierigkeiten

In diesem Unterabschnitt geht es hauptsächlich um die Vorbereitung, also das Preprocessing, der Texte für die verschiedenen Klassifikationsmethoden. Außerdem soll gezeigt werden, wie diese Methoden untereinander und mit im Kapitel 5.5 vorgestellten Verfahren zusammenarbeiten. Dabei soll es hier nicht um eine detaillierte Beschreibung gehen. Diese folgt im nächsten Abschnitt 6.4.

Wir beginnen mit dem noch vollkommen unbearbeiteten Text, der in der Datenbank durch den RSS-Wrapper (siehe Kapitel 8.2) abgespeichert worden ist. Da es hier um Klassifikationsmethoden geht und es darauf ankommt ähnlichen Texten die gleiche Klasse zuzuweisen, sollten wir zunächst die in der Nachricht vorkommenden Wörter stemmen oder lemmatisieren. Denn bei diesen Verfahren zählt die Anzahl des Vorkommens bestimmter Wörter, egal in welcher Deklination oder Konjugation. Das heißt, dass "Aktie" und "Aktien" zwar syntaktisch verschieden sind, für uns allerdings ein und dieselbe Bedeutung darstellen, weil die Grundform der beiden Substantive dieselbe ist. Vereinfacht gesagt, wissen

wir, dass dieser Text in irgendeiner Form von Aktien handelt, und dieser Schluss ist schließlich unabhängig von der Beugung.

Nachdem die wichtigen Wörter - das heißt, die Wörter, die in unserem Wörterbuch stehen (siehe Kapitel 9.4) - so bearbeitet wurden, können wir den Text weiter durch einen Thesaurus normieren. Per Definition ist ein Thesaurus bzw. Wortnetz ein kontrolliertes Vokabular, dessen Begriffe durch Relationen miteinander verbunden sind. Wir benutzen den Thesaurus als ein Netz von Synonymen, um Wörter auf ihr "Grundsynonym" zu reduzieren, analog zum Stemming oder der Lemmatisierung. Wenn wir also zum Beispiel "Ankauf" durch "Kauf" ersetzen, machen wir deutlich, dass es uns nicht auf die Syntax, sondern auf die Semantik des Wortes ankommt. Wir definieren "Ankauf" praktisch als äquivalent zu "Kauf", so wie wir zwischen "Aktie" und "Aktien" keine Unterschiede machen.

Nachdem wir die für uns wichtigen Worte aus dem Text gefiltert und normiert haben, können wir einen Eingabevektor für die verschiedenen in Kapitel 5.5 beschriebenen Verfahren erstellen. Dieser Vektor besteht aus einer festgelegten Anzahl von Komponenten, deren Bedeutung man unterschiedlich wählen kann. In einem Fall könnten es zum Beispiel einzelne Wörter sein, in einem anderen Fall wiederum ganze Klassen. Im letzteren Fall könnte man zum Beispiel "Aktie" und "Optionsschein" einer Komponente zuweisen, deren Bedeutung der Klasse "Wertpapier" entspricht. Außerdem muss entschieden werden, aus welchem Wertebereich die Werte der Komponenten stammen. Auf dies alles wird in der detaillierten Sicht auf die Methoden eingegangen werden.

Der Aufbereitung der Texte zu Vektoren folgt die Anwendung der einzelnen Verfahren, wie sie in Kapitel 6.5 beschrieben sind.

Ergebnis soll eine Einteilung der Nachrichten in negative, positive oder neutrale Nachrichten sein, die dann einen Hinweis auf die Entwicklung des Kursverlaufs geben könnte, also eine Klassifikation der Aktie in under-, market-, oder outperformer ermöglicht.

Noch einmal zusammengefasst:

- Lemmatisierung bzw. Stemming
- Normierung durch Thesaurus
- Erstellung des Eingabevektors für die einzelnen Methoden
- Anwendung einer der unten ausgeführten Methoden
- Automatische Bewertung des Ergebnisses bzw. Einteilung der Aktie zu einer Klasse

5.4 Allgemeine Methoden

5.4.1 Stemming

Mit Grundform- bzw. Normalformreduktion, bezeichnet man eine Gruppe von Verfahren, mit denen morphologische Varianten eines Wortes auf den Wortstamm zurückgeführt werden.

Stemming ist dabei eine recht einfache Form der Wortnormalisierung. Diese Methode zielt auf das Problem der Bearbeitung von unstrukturierten Texten durch einen Rechner, der natürlich keine morphologischen Regeln innerhalb einer Sprache ohne weiteres anwenden kann. Das heißt, er kann ein flexiertes Wort nicht einfach in seine Grundform bringen. Genau dazu wird ein Stemmer benötigt. In [Pai96] wird Stemming folgendermaßen beschrieben:

> *Ein Stemmer sollte die und nur genau die Wortpaare zusammenfassen, die sematisch äquivalent sind und den gleichen Wortstamm teilen.*

So sollte also "Wörter" zu "Wort", "Spiele" zu "Spiel", etc. verändert werden. Schon an diesem Beispiel sehen wir die Schwierigkeiten, die in der deutschen Sprache auftreten. Während es im Englischen recht wenige unregelmäßige Wörter gibt - meist lässt sich ein Verb oder Substantiv im Plural einfach durch abschneiden der Endung -s in den Infinitiv bringen - gibt es in der deutschen Sprache zahlreiche unregelmäßige Wörter.

Als größtes Problem ist allerdings anzusehen, dass Substantive im Deutschen nach dem Geschlecht dekliniert werden, was zur Folge hat, dass kein zuverlässiger oder weitgehend fehlerfreier Stemmer für die deutsche Sprache programmiert werden kann, ohne eine lexikalische Analyse durchzuführen.

Jörg Caumanns hat es dennoch versucht. Sein *GermanStemmer*, der in der Suchmaschine *Lucene - siehe auch [Fou]* eingebaut ist, wurde von der PG genutzt. Der Algorithmus führt im wesentlichen zwei Schritte durch:

- Ersetze einzelne Buchstaben oder Buchstabengruppen im Wort
- Schneide bestimmte Suffixe ab

Die genaue Arbeitsweise des Algorithmus ist dem Paper [Cau] zu entnehmen. Hier sollen nur einige Beispiel gegeben werden:

Eingabe	*Ausgabe*
singt	sing
singen	sing
stören	stö
Mauer	Mau
Kuß	Kuß
Störsender	Stö

Tabelle 5.1: Beispiele für die GermanStemmer-Ausgaben

5.4.2 Lemmatisierung

Die Lemmatisierung hat das gleiche Ziel wie das Stemming. Jedem Wort eines laufenden Textes soll seine Grundform (oder das Lemma) zugeordnet werden. Allerdings werden andere Techniken angewandt, wie zum Beispiel Finite State Transducer oder in Verbindung mit POS-Taggern auch Entscheidungsbäume.

Von der PG wurde der TreeTagger der Universität Stuttgart [IfmS] benutzt. Wie der Name schon sagt, basiert er auf Decision Trees. Der TreeTagger arbeitet mit trigrams, dass heißt, er beachtet die Wortarten der letzten beiden und des aktuellen Wortes. Die Wortart, die die höchste Wahrscheinlichkeit hat - welche in einem Blatt steht - wird dann dem Wort zugewiesen. Für eine genaue Beschreibung siehe [Schb] und [Scha].

Der POS-Tagger gibt als Ausgabe die Wortarten der Wörter und deren Grundformen zurück. Ein kleines Beispiel wäre:

Wort	*Wortart*	*Lemma*
Der	ART	der—die
TreeTagger	NN	TreeTagger
ist	VAFIN	sein
leicht	ADJD	leicht
zu	PTKZU	zu
bedienen	VVINF	bedienen
.	$.	.

Tabelle 5.2: Beispiele für eine TreeTagger-Ausgabe

5.4.3 Thesaurus

Nachdem nun die Nachrichten lemmatisiert sind, besteht der nächste logische Schritt der Verarbeitung darin, Wortgruppen oder Wortpaare, die zwar syntaktisch verschieden, aber semantisch zusammenhängend sind, zu finden und zu sogenannten Synonymgruppen zusammenzufassen. So gehören also zwei oder mehr Wörter zu einer Synonymgruppe, wenn sie in einem bestimmten Kontext die gleiche Bedeutung haben.

Arbeitsintensiv wäre eine manuelle Zusammenfassung von häufig vorkommenden Begriffen innerhalb der Finanznachrichten zu Synonymgruppen. Unser Ansatz ist, einen Thesaurus zu verwenden, der bereits zu vielen Wörtern Synonyme zur Verfügung stellt und diesen gegebenenfalls nach unseren Bedürfnissen zu erweitern.

Die PG hat sich entschieden, den freien offenen deutschen Thesaurus *OpenThesaurus*[1] zu verwenden. Er enthält zur Zeit 35.273 Wörter in 14.375 Synonymgruppen[2]. Ein Beispiel für eine Synonymgruppe wäre die Gruppe 319, bestehend aus:
„Ausbeute, Einnahmen, Erlös, Ertrag, Gewinn, Profit, Rendite, Überschuss“.

OpenThesaurus ist offen, in dem Sinne, dass es registrierten Benutzern möglich ist aktiv am aktuellen Wortschatz mitzuarbeiten. Neue Wörter können erstellt oder bestehende bearbeitet werden. Bei den einzelnen Bedeutungen lassen sich dann z.B. unpassende Wörter löschen oder neue hinzufügen. Der Zugriff auf den Thesaurus (die Abfrage von Synonymen) ist auch für nicht registrierte Benutzer über ein Webinterface möglich, weiter kann ein tagesaktueller SQL-Dump der Datenbank heruntergeladen werden. Anhand eines solchen Dumps wurde der *OpenThesaurus* auch in unsere Datenbank unter *thesaurus* übertragen.

Die wichtigsten Tabellen der Datenbank sind folgende:

- **words:** hier kann für jedes Wort in Spalte *word* die interne ID (*id*) nachgeschlagen werden.

- **word_meanings:** hier kann nun mithilfe der *word_id* eines Wortes, die der *id* aus *words* entspricht, die zugehörige Synonymgruppe (*meaning_id*) des Wortes gefunden werden.

Beispiel einer Anfrage: Ausgabe der Synonymgruppennummer zu ‘Gewinn‘

```
SELECT word_meanings.meaning_id FROM words, word_meanings
WHERE words.id=word_meanings.word_id and words.word=‘Gewinn‘
```

[1] www.openthesaurus.de

[2] Stand: 24.01.2006. Die Anzahl der Wörter ist in den letzen 2 Monaten rasant (um ca. 40%) gewachsen.

5.4.4 Semantik

Semantik oder auch Bedeutungslehre befasst sich mit dem Sinn und der Bedeutung von Sprache beziehungsweise sprachlichen Zeichen. Sie gehört zum Teilgebiet der Linguistik.

Im Kontext von FIPs bedeutet die semantische Analyse von Finanznachrichten eine Übersetzung einzelner Wörter und Sätze in Aussagen, so dass der inhaltliche Sinn solch einer Nachricht durch diese Aussagen erfasst werden kann.

Da wir weder Linguisten sind, noch über ein ausgeprägtes semantisches Wissen über Finanznachrichten verfügen, haben wir uns in dieser PG für die Anwendung von Verfahren aus der Künstlichen Intelligenz (basiert auf symbolischer Darstellung, logische Formeln) und der Computational Intelligence (basiert mehr auf Verarbeitung nummerischer Information, nicht rein zeichenbasiert) entschieden. Die semantische Analyse tritt höchsten in Ansätzen auf, beispielsweise beim POS-Tagging, der Lemmatisierung und der Anwendung eines Thesaurus.

Ohne Frage ist die Betrachtung der Semantik einer Finanznachricht dennoch ein spannendes Teilgebiet, welches sicher bei weitergehenden Betrachtungen zum Einsatz kommen wird.

5.5 Methoden im Detail

5.5.1 Clustering

Vorstellung der Methode

Der k-means clustering Algorithmus gehört zu den unüberwachten Lernalgorithmen. Er ist einer der weit verbreitetsten Clusteringalgorithmen, da er einfach zu implementieren ist, dafür aber im Allgemeinen recht gute Ergebnisse liefert. Sein Ziel ist es, die Daten in k "Bedeutungen" zu unterteilen.

Es wird angenommen, dass die Attribute der Objekte einen Vektorraum formen. Für uns heißt das, dass wir zunächst Dokumentenvektoren erstellen müssen. Ziel ist es nun, die Intra-Cluster Varianz zu minimieren, oder die Funktion:

$$V = \sum_{i=0}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2$$

wobei es k Cluster $S_i, i = 1, 2, \ldots, k$ gibt und μ_i ein Zentroid oder Mittelpunkt der Punkte $x_j \in S_i$ ist.

Der Algorithmus selbst lässt sich schnell beschreiben:

- **Initialisierung**:

  Wähle die k Clusterzentren selbst zufällig oder heuristisch

- **Zuordnung**:
  Ordne jedem Objekt den nächstgelegenen Zentoiden zu
  (unter Angabe einer Distanzfunktion)
- **Zentroidberechnung**:
  Berechne für alle Cluster den neuen, korrigierten Zentroiden
- **Wiederholung**:
  Wiederhole die ersten drei Punkte, bis sich die Zuordnung
  nicht mehr ändert

Der Algorithmus konvergiert, jedoch ist die maximale Anzahl der Iterationen unbekannt. Eine alternative Abbruchbedingung könnte sein, dass wir die maximale Anzahl der Iterationen festlegen.

Anwendung dieser Methode speziell für unser System

Beschrieben werden hier die genutzten Parameter für den Algorithmus.

Für die Distanzfunktion wurde der Hammingabstand gewählt. Die Wörterbuchgrößen unterschieden sich zwischen 4500, 1000 und 22 Wörtern, die Anzahl der Vektorkomponenten korrespondieren mit diesen Zahlen. Einem Wort aus dem Wörterbuch wurde also eine Komponente des Dokumentenvektors zugewiesen. Kommt ein Wort aus dem Wörterbuch im Text vor, wird die entsprechende Komponente des Vektors auf eins gesetzt. Kam ein Wort des Wörterbuches nicht im Text vor, bleibt die Komponente auf null.

Die Initialisierung der Zentroiden geschieht zufällig. Für die Messung der Qualität der Cluster wurde die Varianz benutzt. Enigma-Variationen werden erstellt, indem der Algorithmus auch mit verschiedenen Anzahlen von initialen Zentroiden getestet wird.

Zu den erstellten Dokumentenvektoren ist noch zu sagen, dass sowohl die Wörter des Wörterbuches, wie selbstverständlich auch die in den Texten vorgekommenen Wörter gestemmt wurden.

Anwendbarkeit/Fazit

Leider waren die Ergebnisse alles andere als befriedingend. Oftmals lieferte der Algorithmus nur ein einziges Cluster, was überhaupt keine Aussage mehr über die Nachrichten zulässt. Im besten Fall wurden einige wenige Cluster als Ergebnis erstellt, die aber wiedersprüchliche Nachrichten - dabei handelte es sich nicht einfach nur um einige wenige - bzgl. der Sematik enthielten.

Dennoch wurden einige Lehren aus dem Test des Algorithmus gezogen:

- Je kleiner das Wörterbuch, desto geringere Varianz, d.h. desto höhere Qualität, der Cluster
- Je kleiner das Wörterbuch, desto geringer differenziert sind die Cluster aber auch
- Für differenzierte Aussagen ist Stemming nicht so gut geeignet, da es zu sehr den Sinn verfälscht (siehe Kapitel 6.4.1)
- Im Mittel über zufällige Clusterinitialisierungen bringen gezielte Clustermitten bei der Initialisierung keine wirkliche Verbesserung

Gründe für das schlechte Abschneiden sind:

- Ein schlechtes Wörterbuch (zu wenig Wissen)
- Ausreißer zerstören Cluster
- Struktur der Nachrichten

  Beispiel:

  "Finanzinvestoren interessiert der Gewinn nach Steuern nicht"

  "Interessant ist die gute Gewinnprognose"

  ergeben beide den gleichen Merkmalsvektor.

5.5.2 SVM

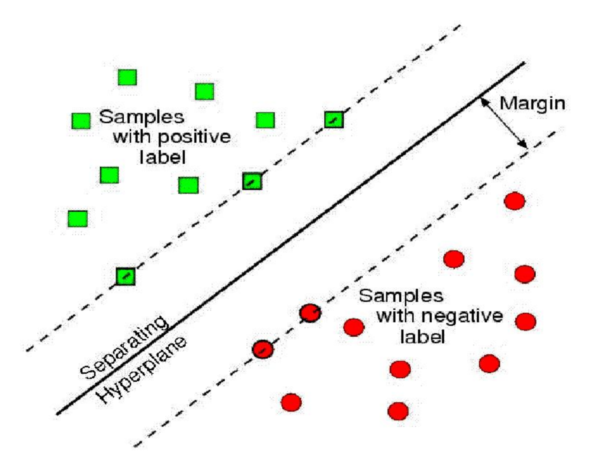


Abbildung 5.1: Beispiel einer optimal trennenden Hyperebene; rote Kreise $\hat{=}$ positive Instanzen, grüne Quadrate $\hat{=}$ negative Instanzen

Vorstellung der Methode

Eine Support-Vektor-Maschine ist ein überwachter Lernalgorithmus, der ein binäres Klassifikationsproblem lösen kann. Die Daten liegen dabei in einem Vektorraummodell vor, d.h. jedes Beispiel wird durch einen Punkt im Vektorraum repräsentiert (siehe Abbildung 5.1 für ein Beispiel im 2-dimensionalen Vektorraum).

Das Ziel der SVM ist die Erstellung einer Hyperebene, die die Mengen der positiven und negativen Beispiele optimal trennt. Eine optimal trennende Hyperebene ist eine Hyperebene mit maximalem Abstand zu den am nächsten liegenden Punkten (siehe Abbildung 5.1).

Diese Hyperebene h wird beschrieben durch einen Gewichtsvektor w und einem Grenzwert w_0:

$$h : w^T x + w_0 = 0$$

und stellt eine Entscheidungsgrenze dar. Für die Klassifikation gilt: wenn das innere Produkt von Gewichtsvektor w und einem Punkt x aus dem Vektorraum grösser als der Grenzwert $-w_0$ ist, dann bekommt dieser Punkt die positive Klasse zugewiesen, ansonsten die negative.

Damit das Verfahren angewendet werden kann, müssen die positiven und negativen Beispielmengen linear separierbar sein. In einem nicht-linear separierbaren Fall wird mittels einer Kernelfunktion φ der n-dimensionale Vektorraum in einen höherdimensionalen Vektorraum transformiert. Durch eine geeignete Kernelfunktion sind die Beispielmengen in diesem neuen Vektorraum immer linear separierbar und die SVM arbeitet dann in diesem hochdimensionalen Vektorraum. Die Entscheidungsgrenze im ursprünglichen Vektorraum kann dadurch beliebig komplex werden.

Auf den ersten Blick könnte dieser hochdimensionale Vektorraum ein Problem für die Berechnung werden. Eine SVM jedoch erstellt die Hyperebene auf der *Basis der Supportvektoren* (Beispiele, die am schlechtesten zu klassifizieren sind) und nicht ausgehend von allen Beispielen. Das Verhältnis $\frac{\#Supportvektoren}{\#Beispiele}$ ist meistens sehr gering, sodass der hochdimensionale Vektorraum keine Probleme bereitet.

Beispiel eines geeigneten Szenarios für SVMs

E-Mail-Klassifikation in Spam und Nicht-Spam Vorgehen hierbei ähnlich wie bei unserem Problem:

- E-Mails bestehen aus Wörtern
- Umwandlung von E-Mails in Vektoren
- Bestimmung von typischen Wörtern für Spam und für Nicht-Spam
- Reihenfolge der Wörter wird ignoriert

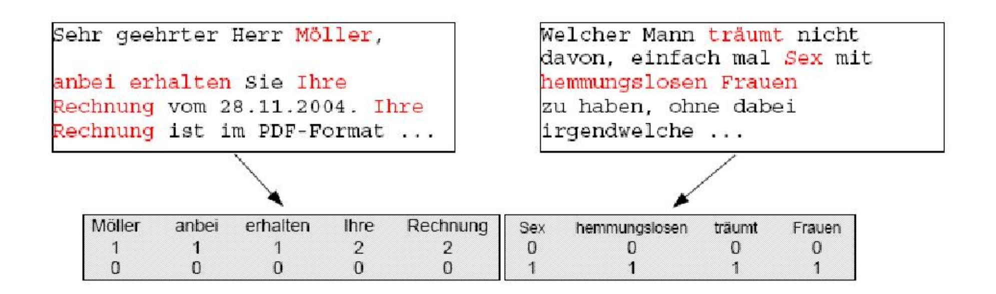


Abbildung 5.2: Beispiel der Problemstellung "Spam-Filterung"; auf der linken Seite ist eine Not-Spam-Nachricht und auf der rechten Seite eine Spam-Nachricht abgebildet. Unten sind die Vektorrepräsentationen zu sehen.

Der Unterschied zwischen unserem Problem und dem Problem der Spam-Filterung ist der, das bei der Spamfilterung der Inhalt von Spam und Nicht-Spam unterschiedlich ist (siehe Abbildung 5.2).

Hierbei konzentriert man sich einfach nur auf die Regel, dass ein Wort besonders aussagekräftig ist, wenn es in vielen Spam-Mails und wenigen Nicht-Spam Mails vorkommt (für Nicht-Spam-Mails analog).

Anwendung dieser Methode speziell für unser System

Problemstellung

Die Finanznachrichten sollten mit Hilfe von SVM so klassifiziert werden, dass man den Benutzern der Finanzinformationsplattform konkrete positive oder negative Nachrichten für eine bestimmte Aktie zurückgeben kann.

Problembehandlung

Für die Klassifikation werde nichtklassifizierte Nachrichten von diversen Newsseiten im Internet in einer Datenbank gespeichert. Von diesen Nachrichten wurden 200 manuell klassifiziert (2 Klassen: "positiv für eine Aktie" und "negativ für eine Aktie"), die dann als Eingabe benutzt werden. Aus den klassifizierten Texten erstellt das *WordVectorTool* [3] die Eingabevektoren im n-dimensionalen Vektorraum (wobei n der Grösse des benutzten Wörterbuches entspricht; $n \approx 13000$).

Des weiteren sind die Features mit dem TFIDF-Maß (*term-frequency inverse-document-frequency*) gewichtet.

Mit Hilfe einer fertigen Implementation einer SVM (*SVMlight*[4]) erzeugen wir unter Verwendung von verschiedenen Kernelfunktionen mehrere Klassifkationsmodelle.

[3]http://www-ai.cs.uni-dortmund.de/SOFTWARE/WVTOOL/index.html
[4]http://svmlight.joachims.org/

Testergebnisse

Die erzeugten Modelle konnten die Fehlerrate auf bestenfalls 27% senken. Eine derart hohe Fehlerrate ist aber für ein Textklassifikationsszenario inakzeptabel.

Mögliche Ursachen für die hohe Fehlerrate

- Keine eindeutige Zuteilung von Worten in eine Klasse (Wort x kann sowohl in einem Text der positiven als auch negativen Klasse vorkommen)
- Nur syntaktische Analyse möglich (bei Finanznachrichten ist aber häufig auch die Semantik für die Analyse wichtig)
- Qualität der Nachrichten (100%-ige Zuordnung zu einer Klasse selten möglich; nicht-relevante Nachrichten konnten nicht immer aussortiert werden)
- Binäre Klassifikationsstruktur möglicherweise nicht ausreichend (vielleicht mehrere Klassen nötig oder die gewählten zwei Klassen waren zu allgemein)

Anwendbarkeit/Fazit

Da bei unserem Problem einzelne Wörter sowohl in der einen als auch anderen Klasse mit einer bestimmten Häufigkeit vorkommen können, ist dieser Ansatz der Klassifizierung mit Hilfe von Support-Vektor-Maschinen eher ungeeignet.

Support-Vektor-Maschinen sind gut geeignet, wenn man die Klassifikation anhand der Häufigkeit der einzelnen Wörter durchnehmen möchte.

Bei unserem Problem allerdings reicht so eine syntaktische Analyse nicht aus, um eine Klassifizierung mit einer akzeptablen Fehlerquote zu erreichen.

5.5.3 ART-2a

Vorstellung der Methode

Beschrieben wird im Folgenden eine Implementierung des ART-2a (adaptive resonance theory-2 advanced) Algorithmus, entwickelt von Stephen Grossberg, Gail A. Carpenter und David B. Rosen.

Allgemein ist ein ART-Netz ein neuronales Netz, das unüberwachtes Lernen unterstützt. Prinzipiell kann jedes Problem gelernt werden, das sich durch einen Vektor codieren lässt. Dieser Vektor wird einer Klasse zugeordnet, die schon gelernt worden ist, oder es wird eine neue Klasse erstellt, falls der Vektor bezüglich eines Ähnlichkeitsmaßes zu verschieden von den bereits erlernten Vektoren ist.

Ein potentieller Vorteil der ART-Netze ist, dass sie das Stabilitäts-Plastizitäts-Dilemma lösen.

- Stabilität bedeutet, dass gelerntes nicht verlernt wird, dass also neue Eingabevektoren die erlernten Gewichte nicht zu stark verändern
- Plastitizät bedeutet, dass das Netz seine Lernfähigkeit beibehält.

ART-Netze lösen das Dilemma, weil sie in einer nicht stationären Welt leben. Während bei anderen Netzen die Anzahl der Klassen schon vor dem Training festgelegt werden muss, kann ein ART-Netz beliebig viele neue Klassen hervorbringen. Wir müssen also nicht unbedingt in Trainings- und Klassifikationsphase unterteilen, in der nichts mehr gelernt wird. Stattdessen wird, wenn ein Eingabevektor nicht in eine schon vorhandene Klasse passt, eine Neue erstellt.

Der ART-2a Algorithmus wurde verwendet, da er gegenüber dem normalen ART-2 Algorithmus Geschwindigkeitsvorteile bringt, dabei allerdings die Ergebnisse praktisch identisch sind, wie die Entwickler zeigen konnten.

Es folgt eine Darstellung der Implementierung des ART-2a Algorithmus:

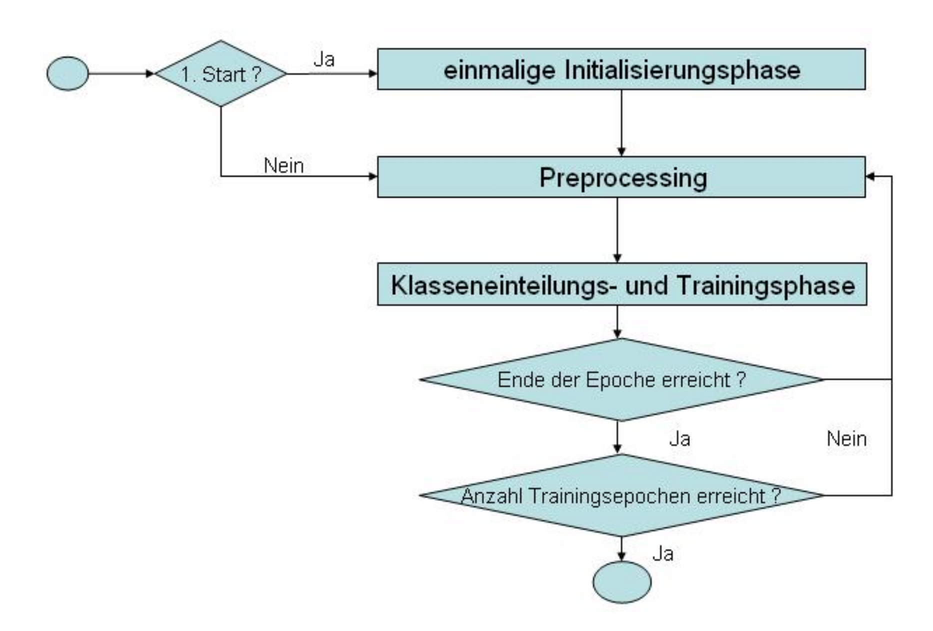


Abbildung 5.3: Übersicht: Implementierung des ART 2a-Algorithmus

In der einmaligen Initialisierungsphase wird die Klassenmatrix, in der die erlernten Klassen als Vektoren gespeichert sind, initialisiert.

In unserer Implementierung ist die Matrix als eine verkettete Liste von Vektoren implementiert, da es überflüssig wäre, Vektoren abzuspeichern, die aber keine Klassen sind, wie es zum Beispiel bei einer Array-Repräsentation der Fall ist.

Anschließend folgt das Preprocessing. Diesem Schritt werden folgende Daten übergeben.

- Eine Eingabematrix, wiederum implementiert als verkettete Liste von Vektoren, die die zu erlernenden/klassifizierenden Vektoren enthält.
- Die Parameter, die die Klassifizierung direkt beeinflussen:
  Aufmerksamkeit: $0 \leq \rho \leq 1$
  Schwellenwert: $0 < \theta < \frac{1}{\sqrt{m}}$, m: Anzahl der Komponenten
  Lernparameter: $0 < \eta \ll 1$
- Ein Parameter α, der die gewünschte Anzahl von Epochen festlegt

Der letzte Parameter legt die Anzahl der Epochen fest, also die Anzahl der Durchläufe durch die Eingabematrix. Ist die Anzahl der Epochen zum Beispiel sechs, wird ein Vektor innerhalb der Eingabematrix sechs mal ausgewählt, die Matrix wird insgesamt sechsmal durchlaufen.

Das Preprocessing lässt sich dann folgendermaßen darstellen:

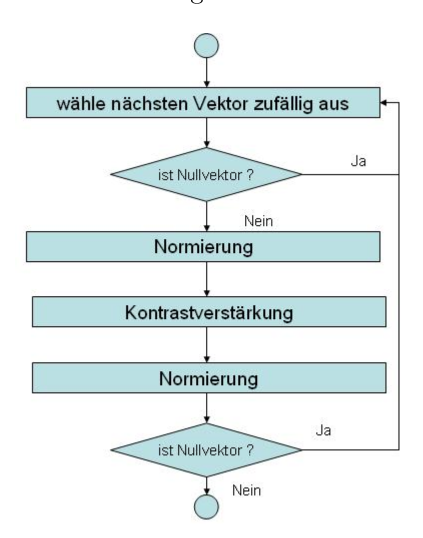


Abbildung 5.4: Übersicht: Preprocessing des ART 2a-Algorithmus

Zunächst wählen wir den nächsten Vektor der Eingabematrix zufällig aus. Dieser Vektor wird dann in dieser Epoche nicht mehr benutzt.

Haben wir einen Nullvektor gewählt, so wird gar nicht mehr weiter verfahren. Wir wählen einfach den nächsten Eingabevektor. Gewöhnlich wird der Nullvektor auf die Nullklasse abgebildet, doch unser Vorgehen ist ja prinzipiell gleich. Wenn wir den Nullvektor sehen, behandeln auch wir ihn gesondert. Es wird nichts mehr gelernt, der Rest der Klassen bleibt unberührt.

Haben wir keinen Nullvektor gewählt, normieren wir ihn zunächst. Nach einer Kontrastverstärkung oder Rauschunterdrückung normieren wir ihn erneut.

Die Rauschunterdrückung lässt sich durch die Implementierung folgender Formel realisieren:

$$v_i = f(x_i) = \begin{cases} x_i & : \quad x_i > \theta \\ 0 & : \quad \text{sonst} \end{cases}$$

Hierbei ist x_i die i-te Komponente des Eingabevektors und v_i die i-te Komponente des neuen Vektors. Durch diese Funktion kann der alte Vektor natürlich wieder auf den Nullvektor abgebildet werden. Wieder wäre es sinnlos, ihn zu erlernen.

Nun befinden wir uns in der Klasseneinteilungs- und Trainingsphase. Wie oben schon angedeutet sind dies nicht zwei gänzlich voneinander getrennte Phasen. Stattdessen ist die Trainingsphase von der Klasseneinteilungsphase abhängig, was auch die Besonderheit dieser Netze ausmacht.

Zunächst versuchen wir einen gegebenen Eingabevektor zu klassifizieren. Gelingt uns dies, geben wir einfach die Klasse aus und passen die Gewichte dem klassifizierten Vektor an. Wie weit wir ihn anpassen, hängt vom Lernparameter ab, dazu später mehr.

Gelingt uns jedoch keine Klassifikation, erstellen wir eine neue Klasse. Die Gewichte dieser neuen Klasse entsprechen genau den Komponenten des Eingabevektors.

Eine detaillierte Beschreibung der Implementierung sieht folgendermaßen aus:

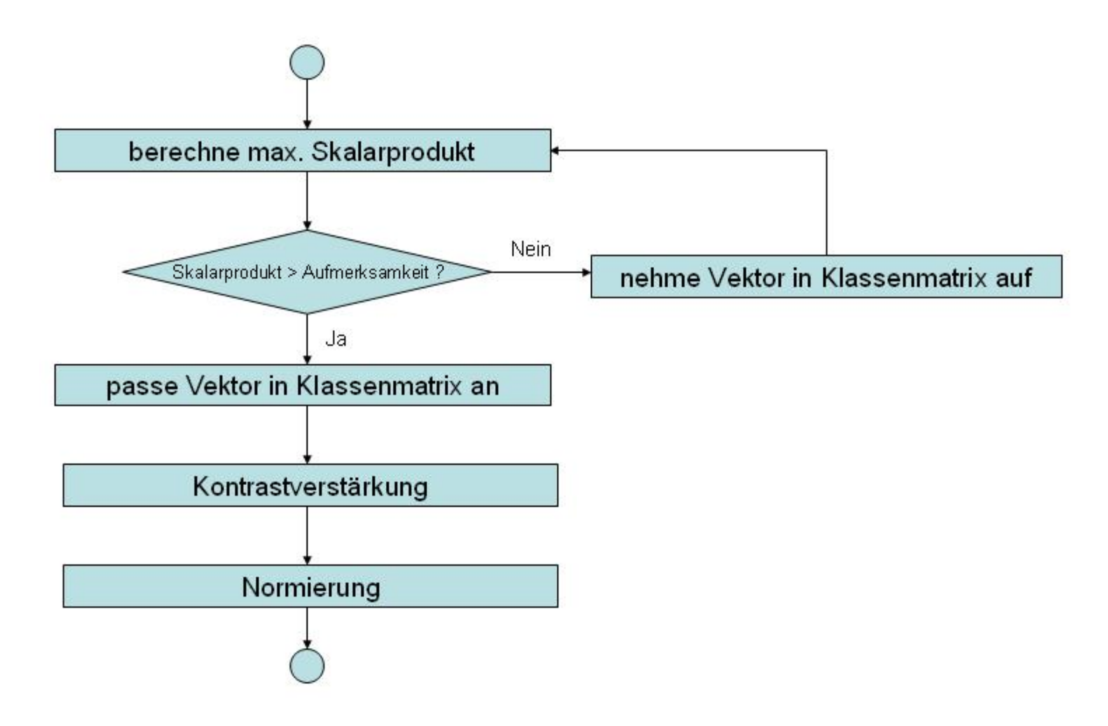


Abbildung 5.5: Übersicht: Klasseneinteilung

Zum ersten Schritt: Das Skalarprodukt der Vektoren der Klassenmatrix mit dem Eingabevektor wird ermittelt und davon anschließend das Maximum genommen. Weil es sich hier immer um Einheitsvektoren handelt, gilt folgende Formel:

$$\begin{aligned} T_J &= \max\{\overline{v} \cdot \overline{w_j} : \text{für alle Klassen j}\} \\ &= \max\{|v_i||w_{ij}|\cos(\phi) : \text{für alle Klassen j und allen Komponenten i}\} \\ &= \max\{\cos(\phi_i) : \text{über allen Klassenvektoren j und dem Eingabevektor}\} \end{aligned}$$

Das heißt, dass wir den Cosinus der Winkel zwischen den Klassenvektoren und dem Eingabevektor berechnen. Je größer der Cosinus, umso kleiner die Winkel, umso ähnlicher die Vektoren. Mit anderen Worten: Als Ähnlichkeitsmaß wird der Cosinus benutzt.

Im nächsten Schritt wird dieser Wert mit dem Aufmerksamkeitsparameter verglichen: Also bezeichnet der Ähnlichkeitsparameter eigentlich nur den Cosinuswert, ab wann wir den Eingabevektor als genügend ähnlich zu einer Klasse akzeptieren.

Falls uns die Ähnlichkeit des Eingabevektors nicht überzeugt, wird eine neue

Klasse erzeugt und der Vektor in die Klassenmatrix aufgenommen.

Andernfalls geben wir die Klasse aus und passen die Gewichte dieser Klasse nur an. Hier spielt der Lernparameter eine wesentliche Rolle:

$$\bar{t} = \eta\overline{v} + (1-\eta)\overline{w_J^{old}}$$

$$\overline{w^{new}} = \frac{\bar{t}}{|\bar{t}|}$$

$\overline{w^{old}}$ bezeichnet den alten, $\overline{w^{new}}$ den neuen Klassenvektor, $\overline{v}$ den Eingabevektor. Er ist schon im Preprocessing normiert worden. Nachdem die Formel berechnet worden ist, bildet der kontrastverstärkte, normierte Vektor den neuen Klassenvektor.

Rauschunterdrückung ist hier noch einmal angewendet worden, weil $\overline{w^{new}}$ unter Umständen sehr kleine Komponentenwerte beinhalten kann, die aber praktisch keine Rolle spielen.

Anwendung dieser Methode speziell für unser System

Unser Ziel ist es, die Nachrichten in drei Klassen einzuteilen: Positive, negative oder neutrale Bedeutung für eine Unternehmung bzw. Aktie. Dazu werden die Komponenten des Dokumentenvektors selbst als Klassen gesehen. So wird jeder einzelnen Komponente eines Vektors folgende Klassen zugeordnet:

- Wertpapier
- Analyse
- Politik
- Börse
- ...

Dies ist natürlich nur ein kleiner Auschnitt des Vektors, der tatsächlich benutzt wurde. Der originale Vektor hat 19 Komponenten.

Den einzelnen Komponenten werden dann durch das Preprocessing einzelne Wörter zugewiesen. Zum Beispiel für die Kategorie Börse:

- Börse
- DAX
- MDAX

- Terminbörse
- Wertpapiermarkt
- ...

Das bedeutet, dass das Wort "Terminbörse" oder "Wertpapiermarkt" die gleiche Bedeutung wie das Wort "Börse" hat, da ihnen dieselbe Komponente zugewiesen ist.

Mit anderen Worten haben wir hier also nicht nur den Thesaurus benutzt, sondern ihn gleich noch ein wenig erweitert, indem wir dieser Komponente ja auch das Wort "MDAX" zugewiesen haben. Aus unserer Sicht war dies sinnvoll, da wir zunächst nur sehen wollten, ob sich - branchenbedingt - sowieso schon ähnliche Texte noch mit den klassischen, einfachen Methoden der Textklassifikation unterscheiden lassen. Dann wollten wir einen Schritt weitergehen, um zu sehen, ob wir die Nachrichten tatsächlich in die drei Kategorien einteilen können.

Anwendbarkeit/Fazit

Für den eigentlichen Zweck ist diese Methode nicht zu gebrauchen. Dies wurde schon bei der Konstruktion der Vektoren klar. Denn ohne jegliche Anwendung von Semantik implementieren wir - bezogen auf die Aufgabe - ja nur eine recht schwierige Abzählmethode. Die Wörter werden einzeln, also zusammenhangslos, betrachtet. Dadurch können aus Nachrichten, die von verschiedenen Unternehmungen handeln, unmöglich Tendenzen für ein einzelnes Unternehmen bestimmt werden. Außerdem macht ein kleines Wörtchen wie "nicht" doch schon einen deutlichen Bedeutungsunterschied aus. Es kommt allerdings darauf an, WO es im Satz steht. Dies verhält sich natürlich nicht nur für Adverbien so, sondern auch für Adjektive und Verben.

Kurz gesagt, kommen wir hiermit zu keinen feinen Unterscheidungen der Semantik der Texte.

Ein positives Ergebnis ist allerdings auch zu vermelden: Die Klassifikation von Dokumenten aus gleichen oder ähnlichen Branchen funktioniert mit dieser Methode. Es kommt nur darauf an, den Dokumentenvektor einigermaßen geschickt zu konstruieren.

5.5.4 Entscheidungsbäume

Entscheidungsbäume werden bei überwachter Klassifikation eingesetzt. Dabei sollen Objekte, die durch Attribut/Werte-Paare beschrieben sind, in bestimmte, vorher festgelegte Klassen, einsortiert werde.
Ein Teil der Objektmenge wird zur Trainingsmenge, aus denen der Entscheidungsbaum aufgebaut wird. Die übrigen Objekte werden dann mit Hilfe des

Entscheidungsbaumes in die jeweiligen Klassen eingeteilt. Meist werden binäre Entscheidungsbäume eingesetzt, wobei bei jedem Attribut entschieden wird, ob es für das Objekt zutrifft oder nicht.

Entscheidungsbäume sind aufgebaut aus Knoten und gerichteten Kanten. Die Wurzel und die inneren Knoten repräsentieren Attribute und die Kanten Attributbelegungen. Die Blätter repräsentieren die Klassen, in welche die Objekte einsortiert werden. Diese Objekte werden am Ende eines vollständigen Pfades klassifiziert. An jedem inneren Knoten wird die Kante gewählt, bei der die Attributbelegung der Kante mit der des Objektes übereinstimmt. Um einen Entscheidungsbaum möglichst klein zu halten, um die Objekte recht effektiv einzuteilen, ist die Wahl des Attributes, welches als nächstes betrachtet wird, von großer Wichtigkeit. Es wird immer das Attribut gewählt, welches den größten Informationsgehalt besitzt. Das ist dasjenige, das die meisten Objekte klassifiziert.

Erster Ansatz

Das Hauptproblem bei der Anwendung dieser Methode für unser System besteht darin, dass wir keine vernünftigen Attribute finden, nach denen die Texte einzuteilen sind. Deshalb haben wir diese Methode ein wenig abgewandelt. Aus einem Entscheidungsbaum liest man im Endeffekt eine Argumentationskette in der folgenden Form ab: 'IF Attributwert#1 des Objektes = Attributwert#1 der Kanten UND ... UND Attributwert#n des Objektes = Attributwert#n der Kanten THEN Objekt gehört in Klasse m'. Wir haben nun als Attribute einzelne Wörter verwendet und daraus Argumentationsketten aufgebaut, die dann in die Klassen 'Positiver Text' bzw. 'Negativer Text' eingeteilt werden. 'Positiver/Negativer Text bedeutet hier, dass der Inhalt des Textes für ein Unternehmen eher positiv oder negativ ist. Diese Einteilung soll einen Hinweis auf den Aktienverlauf geben; ob nämlich das Unternehmen under-, markt- oder outperformer ist. Wir haben ein gestemmtes Wörterbuch auf Grundlage der Trainingstexte erstellt und uns Wörterkombinationen überlegt, die, wenn sie zusammen in einem Text auftauchen, auf einen positiven oder negativen Text schließen lassen. Je nachdem, nach wie vielen Wörtern in einem Text gesucht wird, erhält man für eine 'positive' Argumentationskette +1, +2 oder +3 (bei drei oder mehreren Wörtern) Punkte und für eine 'negative' Argumentationskette -1, -2 oder -3 Punkte. Ketten, die aus mehreren Wörtern bestehen, werden also stärker gewichtet. Am Ende wird die Summe der Punkte aller Argumentationsketten gebildet. Ist die Summe negativ, wird der Text als 'Negativer Text' eingeteilt, ist sie positiv, als 'Positiver Text'. Wenn die Summe = 0 ist, wird der Text als 'Neutraler Text' eingestuft. Auch dann, wenn überhaupt keine Argumentationskette auf den Text zutrifft.

Auch die abgewandelte Version der Entscheidungsbäume ist nicht gut einsetzbar. Das liegt in erster Linie wohl daran, dass unser Wörterbuch nicht vollständig ist und wir deshalb auch nicht ausreichend viele Wörterkombinationen erstellen können, um die Texte vernünftig zu klassifizieren.

Zudem wird die Semantik des Textes nicht berücksichtigt und bspw. wird

nicht zwischen 'Insolvenz' und 'keine Insolvenz' unterschieden, weil wir Stopp-Wörter nicht in unserem Wörterbuch haben. 'Keine' kann ja auch ganz woanders im Text stehen und inhaltlich nichts mit Insolvenz zu tun haben. Dieser Zusammenhang ist nicht erkennbar, da nur binär nach Vorkommen oder Nichtvorkommen im Text gesucht wird. In den zwei folgenden Abbildungen sind die Testergebnisse aufgeführt.

Vom System bewertet		von uns bewertet	Abweichung
Nachricht Nr. 0 wird vom System folgendermassen bewertet: —		—	0
Nachricht Nr. 1 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 2 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 3 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 4 wird vom System folgendermassen bewertet: +++		ooo	1
Nachricht Nr. 5 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 6 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 7 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 8 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 9 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 10 wird vom System folgendermassen bewertet: ooo		+++	1
Nachricht Nr. 11 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 12 wird vom System folgendermassen bewertet: +++		—	2
Nachricht Nr. 13 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 14 wird vom System folgendermassen bewertet: ooo		+++	1
Nachricht Nr. 15 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 16 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 17 wird vom System folgendermassen bewertet: ooo		+++	1
Nachricht Nr. 18 wird vom System folgendermassen bewertet: +++		—	2
Nachricht Nr. 19 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 20 wird vom System folgendermassen bewertet: —		—	0
—bis hier her waren es erlernte Beispiele———			8

Abbildung 5.6: Entscheidungsbaum Tests mit den gelernten Beispieltexten

Zweiter Ansatz

Der erste Ansatz hat zwei Schwachstellen, die wir in einem neuen Ansatz zu verbessern verucht haben. Zum einen wurden im ersten Ansatz die Vorteile der Entscheidungsbaum-Theorie nicht vollständig ausgenutzt. Das lag daran, dass wir selber manuell festgelegten, bei welcher Wortkombination eine Nachricht als gut oder als schlecht zu klassifizieren ist. Dies hat zum einen den Nachteil, dass für die Klassifikation der Nachrichten nur eine begrenzte Anzahl von Kriterien festgelegt werden konnten. Zum anderen können Nachrichten positiv oder negativ sein aus Gründen, die wir übersehen haben.

Die zweite wichtige Schwachstelle des ersten Ansatzes ist, dass er keine Lernkomponente besitzt. Anstatt also die Kriterien für die Güte einer Nachricht selber festzulegen, ist es klüger, anhand vieler bewerteter Nachrichten, systematisch herauszufiltern, was die Kriterien für die Güte der Nachrichten sind.

Aus diesen Gründen ergibt sich die Notwendigkeit, einen an unser Problem angepassten, augereiften Entscheidungsbaum-Algorithmus zu verwenden. Nach einer Recherche stellte sich der als Teil des Systems WEKA implementierte ID3-Entscheidungsbaum-Algorithmus als geeignet heraus. WEKA steht für Waikato

Vom System bewertet		von uns bewertet	Abweichung
Nachricht Nr. 21 wird vom System folgendermassen bewertet: —		+++	2
Nachricht Nr. 22 wird vom System folgendermassen bewertet: +++		ooo	1
Nachricht Nr. 23 wird vom System folgendermassen bewertet: ooo		+++	1
Nachricht Nr. 24 wird vom System folgendermassen bewertet: ooo		+++	1
Nachricht Nr. 25 wird vom System folgendermassen bewertet: ooo		+++	1
Nachricht Nr. 26 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 27 wird vom System folgendermassen bewertet: +++		ooo	1
Nachricht Nr. 28 wird vom System folgendermassen bewertet: +++		ooo	1
Nachricht Nr. 29 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 30 wird vom System folgendermassen bewertet: +++		ooo	1
Nachricht Nr. 31 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 32 wird vom System folgendermassen bewertet: +++		ooo	1
Nachricht Nr. 33 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 34 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 35 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 36 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 37 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 38 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 39 wird vom System folgendermassen bewertet: +++		+++	0
Nachricht Nr. 40 wird vom System folgendermassen bewertet: +++		+++	0
			10

Abbildung 5.7: Entscheidungsbaum Tests mit den neugelernten Texten

Environment for Knowledge Analysis und wurde an der University of Waikato in Neuseeland entwickelt.

Um diesen Algorithmus für unser Problem zu testen, bedienten wir uns einem neuen Wörterbuch mit 302 Wörtern. Wir untersuchten und bewerteten ca. 200 neue Finanznachrichten. Aus diesen Nachrichten suchten wir uns 68 Nachrichten heraus, die unserer Meinung nach klar positiv oder klar negativ waren. Aus diesen Nachrichten wurden boolsche Vektoren der Länge 302 generiert. Wenn ein Wort aus dem Wörterbuch in einer Nachricht vorkam, wurde die entsprechende Stelle des Nachrichtenvektors auf 1 gesetzt. 30 Nachrichtenvektoren wurden als Trainingsbeispiele verwendet, wobei diese aus 15 positiven und 15 negativen bestanden. Der Rest der Vektoren wurden als Test-Beispiele verwendet. Den ID3-Entscheidungsbaum passten wir so an, dass er die Nachrichten in nur zwie Klassen einteilte: gut und schlecht. In drei Durchläufen mit jeweils verschiedenen Trainings- und Testbeispielen wurde der Entscheidungsbaum mit den Trainingsdaten trainiert und die Testdaten auf dem Entscheidungsbaum getestet. Das Ergebnis war, dass unser Entscheidungsbaum bei allen Durchläufen knapp über 50 Prozent der Testnachrichten richtig klassifizierte.

Dieses Ergebnis hat aber leider keine große Aussagekraft, da Finanznachrichten inhaltlich auf verschiedenste Weise aufgebaut sein können und wir deswegen nicht davon ausgehen können, dass wir alle wichtigen Wörter, die entscheidend für die Klassfikation einer Nachricht sind, in unserem Wörterbuch erfasst haben. Hinzu kommt, dass hier wie beim ersten Ansatz die Semantik und Reihenfolge des Auftretens der Wörter nicht berücksichtigt wird.

Beispiel	Sport	Art	Ort	Ebene	Tag	Anschauen
X1	Fußball	Mannschaft	draußen	national	Samstag	+
X2	Hockey	Mannschaft	draußen	national	Samstag	+
X3	Bodenturnen	Einzel	drinnen	welt	Samstag	-
X4	Handball	Mannschaft	drinnen	national	Samstag	+
X5	Zehnkampf	einzel	draußen	welt	Sonntag	-

Tabelle 5.3: Trainingsbeispiele für Konzeptlernen

5.5.5 Konzeptlernen

Vorstellung der Methode

Ein Konzept ist eine einstellige Funktion $c : M \rightarrow \{0, 1\}$.
M ist hierbei die Grundmenge von Beispielen, anhand derer das Konzept c gelernt werden soll. Ein Konzept wird mit Hilfe von positiven und negativen Beispielen trainiert, d.h. der Lernvorgang erfolgt durch explizite Angabe von Beispielen, die akzeptiert oder nicht akzeptiert werden sollen. Für positive Beispiele soll gelten: c(X) = + , für negative c(X) = -.

Zur besseren Veranschaulichung soll zunächst dieses Beispiel aus der DVEW-Vorlesung dienen:

Beispiel:

Zu erlernendes Konzept: Sportsendungen, die Paul schaut.

Die Beispiele der Grundmenge M werden als Tupel von Attributen in der Beispielsprache formuliert. Hier: (Sport, Art, Ort, Ebene, Tag)

Menge an Trainingsbeispielen:

Die *Konzeptsprache* unterscheidet sich von der *Beispielsprache* insofern, als dass hier noch die Sonderzeichen ,,?” und ,,-” erlaubt sind.

- **?** : kennzeichnet, dass jeder Attributwert für dieses Attribut erlaubt ist
- **-** : kennzeichnet, dass kein Attributwert für dieses Attribut erlaubt ist

Eines der *Zielkonzepte*, das anhand dieser Grunmenge gelernt werden soll, ist:

(?, Mannschaft, ?, national, Samstag)

Anwendung dieser Methode speziell für unser System

Die **Probleme**, die sich hierbei für unsere Nachrichtenklassifikation ergeben, sind folgende:

Man benötigt also für jede Klasse ein Konzept, welches bestimmt, ob eine Nachricht zu der Klasse gehört oder nicht.

1. Problem: Das Konzept (für jede Klasse) muss bekannt sein. Das ist bei uns bisher nicht der Fall.
2. Problem: Für jedes Konzept müssen geeignete Attribute gefunden werden, die dann jede Nachricht der Klasse aufweisen muss. Das wird wahrscheinlich nur anhand einer semantischen Analyse der Texte funktionieren. Dazu haben wir bisher keine Ergebnisse.
3. Problem: Für jedes Konzept müssen positive und negative Beispiele vorhanden sein, die wir von Hand klassifizieren. Da wir nicht entscheiden können, welches Attribut den Ausschlag gibt, ist das nicht möglich.

Zwischen-Fazit:

Das Konzeptlernen eignet sich nach dieser Voranalyse kaum zur Textklassifizierung. Hauptgrund sind uns fehlende Erkenntnisse über semantische Abhängigkeiten in den Texten, mittels denen die Attribute geeignet ,,gruppiert" werden könnten, um die volle Mächtigkeit des Konzeplernens auszuschöpfen. Weiterhin muss für jedes zu erlernende Konzept eine Menge an Beipielen zum ,,Training" des Lernsystems von Hand bewertet werden, was allerdings auch Chancen bietet, dass das System die Semantik u.U. selbst lernen kann, wenn dieses von unseren Vorgaben lernt.

Abstraktion → Implementierungsansatz

Damit das Konzeptlernen überhaupt mit der uns zugrundeliegenden Aufgabenstellung der Textklassifizierung (= *ohne explizite Semantikanalyse, demnach nur mit Wissen über Syntax und syntaktische Abhängigkeiten*) ,,einigermaßen sinnvoll" lauffähig gemacht werden kann, müssen die vorher genannten Eigenschaften und Eingaben des Lernverfahrens **geeignet abstrahiert werden**:

- Zunächst wird mit sehr wenigen Konzept-Klassen begonnen, da alle Beispiele für jedes Konzept ,,von Hand" bewertet werden müssen und wir außerdem prinzipiell gar nicht wissen, welche sinnige Klasseneinteilungen bestehen könnten.
  Eines ist aber klar: Die Konzepte
  - ,,Nachricht bestätigt Outperformer"
  - ,,Nachricht bestätigt Marketperformer"

  - „Nachricht bestätigt Underperformer"

  sind essentiell für unsere Bewertung und somit seien diese Konzepte die Grundlage einer ersten Implementierung. Im Folgenden wird sich zunächst auf das Konzept des Outperformers konzentriert in der Hoffnung, wenigstens Nachrichten klassifizieren zu können, die ein Unternehmen als Outperformer bestätigen würden. Die anderen Konzepte werden dann analog trainiert (mit ggf. anderen Beispielen und anderen Bewertungen)

- Es wird anhand von syntaktischen Abhängigkeiten gelernt *(Semantikerkenntnisse liegen nicht vor)*

  → **Wörterbuch wird benötigt**

- Die Attribute der Beispielsprache sind nun nicht mehr mehrwertig, sondern binär.

  → Jedes Attribut steht für ein Wort
  → Jeder Attributwert ist entweder 0 oder 1 (Wort kommt im Text vor (1) oder nicht(0))

Zur weiteren Verbesserung wird ein Stemming des Wörterbuches betrieben.

Jeder vorliegende Nachrichten-Text kann nun zu einem Beispiel als Eingabe konvertiert werden, wenn er von uns manuell bewertet wurde, das Lernverfahren also weiß, ob es den Text für das aktuelle Konzept (Outperformer) akzeptieren oder nicht akzeptieren soll.

Implementierung

Das von uns verwendete Verfahren zum Training des Systems ist das „Versionenraumverfahren", welches sich anbietet, da es in zwei Richtungen lernt (speziellste und allgemeinste Hypothesen).

Die Beispiele (Nachrichtentexte) werden in einer Textdatei gespeichert, getrennt von einem Trennwort und mit Angabe, ob das Beispiel akzeptiert werden soll oder nicht (1 oder 0). Das Lernverfahren liest nun jedes der vorliegenden Beispiele mit unserer Bewertung ein und trainiert somit seine Entscheidungsfindung (passt die speziellsten und allgemeinsten Hypothesen an). Nach dem Training mit den vorliegenden Beispielen ist das Konzept gelernt und wir können prüfen, ob das System nun weitere Beispiele (Nachrichtentexte) genau wie wir als Outperformer bestätigen würde oder ob die Ergebnisse des Systems unsinnig sind.

Anwendbarkeit/Fazit

Testen des Lern-Systems Konzeptlernen:

Zunächst: Was könnte die Leistung des Konzeptlernens beeinflussen?

Folgende Parameter könnten die Leistung beeinträchtigen:

- **Wörterbuchgröße**
  (zu groß, zu klein)
- **Zu wenig Trainings-Beispiele**
  - zu wenig positive Beispiele
  - zu wenig negative Beispiele

  (System hat nicht genug Wissen, um nachfolgende Beispiele vernünftig zu klassifizieren)
- **Falsche Bewertung (bestätigt Outperformer (positives Beispiel) oder nicht)**
  (da wir selbst die Bewertung dem System mitteilen, versuchen wir Fehler soweit wie möglich zu vermeiden, und aussagekräftige Trainingsbeispiele zu finden. Somit wäre dieser Punkt zu vernachlässigen.)

Wenn trotz Variation dieser Parameter das System keine sinnigen Klassifizierungen machen kann, dann müssen wir davon ausgehen, dass das Konzeptlernen (leider) keine geeignete Klassifikationsmethode zum Einteilen von Nachrichtentexten darstellt.

Probelauf 1 (sehr kleines Wörterbuch)

Trainings-Parameter:

Wörterbuchgröße : 126
Anzahl Beispiele insgesamt : 20
davon positive : 13
davon negative : 7

Test-Parameter:

Anzahl Beispiele insgesamt : -
davon positive : ?
positiv richtig klassifizierte : ?
davon negative : ?
negativ richtig klassifizierte : ?

Auswertung(1):

Das System konnte nicht getestet werden, da bereits nach dem Lernen des 10. Beispiels der Versionenraum inkonsistent geworden ist. Eine Lösungsmöglichkeit für dieses Problem ist ein größeres Wörterbuch. (s. Auszug aus Laufergebnisse):

fertig mit Lernen von Beispiel 9
Versionenraum sieht nun so aus :
0 . speziellste Hypothese: < ... >
0 . allgemeinste Hypothese: < ... >
Ist Versionenraum konsistent ? : true

Lerne grade das BSP 10 : (positiv) < ... >
Versionenraum vorher :
0 . speziellste Hypothese: < ... >
0 . allgemeinste Hypothese: < ... >
fertig mit Lernen von Beispiel 10
Versionenraum sieht nun so aus :
0 . speziellste Hypothese: < ... >
Ist Versionenraum konsistent ? : false

Probelauf 2 (sehr großes Wörterbuch)

Trainings-Parameter:

Wörterbuchgröße : 13000
Anzahl Beispiele insgesamt : 20

davon positive : 13
davon negative : 7

Test-Parameter:

Anzahl Beispiele insgesamt : -

davon positive : ?
positiv richtig klassifizierte : ?
davon negative : ?
negativ richtig klassifizierte : ?

Auswertung(2):

Es traten immense Speicherprobleme (out of memory exception) bei Wörterbüchern mit mehr als 2500 Wörtern auf, was vor allem daran liegt, dass nach der Implementierung eine Hypothese dann bereits auch 13000 Int-Werte beinhaltet (nicht anders zu lösen). Wenn nun bei Anpassung des Versionenraumes auch mehrere Hypothesen angelegt und verwaltet werden müssen, reicht der lokale Speicher nicht mehr aus um alle diese Daten aufzunehmen.

Probelauf 3 (maximal mögliches Wörterbuch)

Trainings-Parameter:

Wörterbuchgröße : 2500
Anzahl Beispiele insgesamt : 20

davon positive : 13
davon negative : 7

Test-Parameter:

Anzahl Beispiele insgesamt : ?

davon positive : ?
positiv richtig klassifizierte : ?
davon negative : ?
negativ richtig klassifizierte : ?

Auswertung(3):

Bei dieser Wörterbuchgröße, welche das maximal zu verarbeitende Potential voll ausschöpft, wurde der Versionenraum auch wieder inkonsistent. Auch bei Varianten zwischen 126 und 2500 Wörtern wurde unser Versionenraum bereits nach sehr wenigen Beispielen inkonsistent.
Da es nun nicht daran liegen kann, dass wir zu viele Beispiele verwendet haben (im Gegenteil, es müssten erheblich mehr Beispiele zur Betrachtung herangezogen werden um vernünftige Ergebnisse überhaupt erzielen zu können), lässt sich nun feststellen, dass das Konzeptlernen nicht geeignet für die Klassifikation von Finanznachrichten anhand syntaktischer Gesichtspunkte ist.

FAZIT:

Das Konzeptlernen ist nicht ausdrucksstark genug, um Finanznachrichten klassifizieren zu können. Schon nach dem Lernen von wenigen Beispielen tritt eine Inkonsistenz des Versionenraums auf. Die Beispiele lassen sich also nicht ausreichend differenzieren, so dass der Algorithmus nach kurzer Zeit bei gegebener Ausgangslage nicht mehr zwischen positiven und negativen Beispielen unterscheiden kann. Das kann zum einen wie oben angedeutet an der Anzahl verschiedener Attribute (sprich: Anzahl der Wörter), oder aber an der Ausdrucksstärke der Attribute liegen.

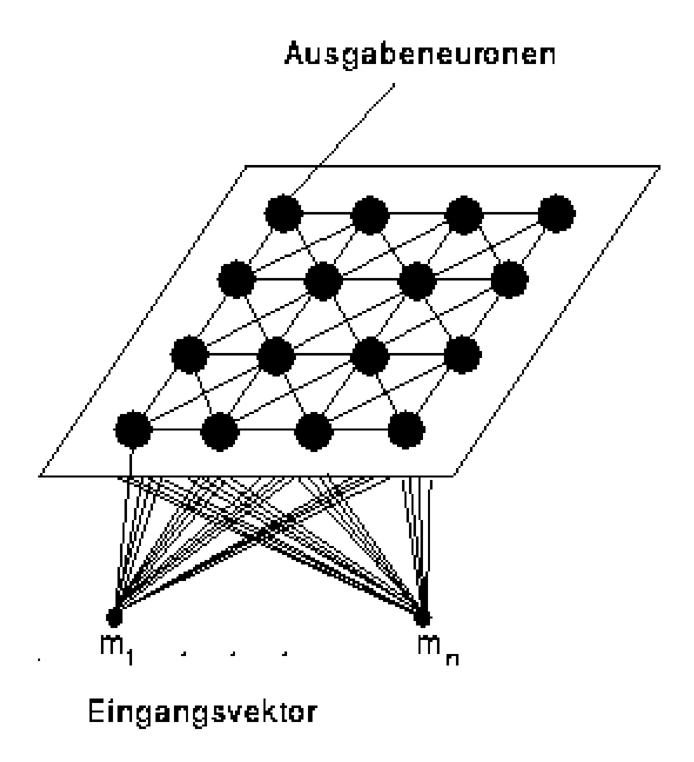


Abbildung 5.8: Self organizing map

5.5.6 SOMs

Vorstellung der Methode

Die SOM (self organizing map, Abbildung 5.8) ist ein spezielles neuronales Netz mit einem unüberwachten Lernverfahren. Die SOM ist eine Karte die aus einer Reihe von Knoten, welche in einer Gitterstruktur, durch die eine Nachbarschaftsbeziehung zwischen den Knoten definiert ist, angeordnet sind, besteht. Jeder dieser Knoten besitzt einen zufällig initialisierten Gewichtsvektor (später eine Gewichtsmatrix) von der Dimensionalität des Eingaberaums. Während des Trainingsprozesses werden die Eingabevektoren (später Eingabematrizen) in zufälliger Reihenfolge dem Netzwerk präsentiert. Für jeden Knoten wird die Aktivierung entsprechend einer Aktivierungsfunktion (Euklidische Distanz) berechnet und jener Knoten mit der höchsten Aktivierung (d.h. geringste Euklidische Distanz) als Gewinner ausgewählt. Im nächsten Schritt werden nun die Gewichtsvektoren des Gewinners und seiner Nachbarknoten innerhalb der Netztopologie einer monoton fallenden Lernrate folgend dahingehend verändert, daß sie dem angelegten Eingabevektor ähnlicher werden. Dies führt letztendlich dazu, daß benachbarte Vektoren aus dem Eingaberaum auf benachbarte Knoten abgebildet werden, d.h. es entsteht eine topologieerhaltende Abbildung.

Die Idee eine SOM zu verwenden um Nachrichten zu klassifizieren soll zu dem erhoffte Ziel führen, dass ähnlich Nachrichten auch topologisch in der Gitterstruktur nahe bei einander sind. Das Problem besteht auch bei diesem Verfahren darin, eine geeignete Repräsentation einer Nachricht zu finden, insbesondere eine Repräsentation, die die Syntax und die Semantik berücksichtigt. Es werden im folgenden nur Nachrichten repräsentiert, die durch den RSS-Wrapper aus dem Internet geladen wurden, die den SPAM-Filter passiert haben und denen ein Unternehmen aus dem DAX30 zugeordnet werden konnte. Die Nachricht muss als Synonymgruppenvektor vorliegen. Das bedeutet, dass die Nachricht nur noch aus IDs von Synonymgruppen des OpenThesaurus besteht und keine Stoppwörter mehr enthält. Eine Synonymgruppenvektor einer Nachricht hat beispielsweise folgende Repräsentation:

- Nachricht: Der Chipkonzern Infineon ist tief in die Verlustzone zurückgefallen.
- Synonymgruppenvektor: $< (NN), (NN), (VAFIN)7530/1626...$

Zur Herleitung dieses Repräsentanten siehe dazu die Kapitel Lemmatisierung, Stemming und Thesaurus. Bei diesem Vektor geht allerdings die Semantik der Nachricht verloren. Was bis jetzt erreicht worden ist, ist eine Maximierung der Allgemeinheit einer Nachricht. Da ähnliche Wörter (mit gleichem Sinn) in Gruppen zusammengefasst worden sind und die einzelnen Wörter vorher alle auf ihre Grundform reduziert worden sind.

Um den Zusammenhang zwischen Wörtern zu repräsentieren ist eine Erweiterung des Synonymgruppenvektors auf eine Synonymgruppenmatrix notwendig. Die Elemente der Matrix enthalten Wahrscheinlichkeiten, die informell bedeutet, wie wahrscheinlich es ist, dass ein Wort X in einem Satz auftaucht, wenn vor Wort X das Wort Y im Satz vorkam. Bevor dies formalisiert werden kann, gilt es noch als Vorbedingung ein feststehendes Wörterbuch vorauszusetzen. Wie oben die Nachricht schon zeigt, haben alle Nachrichten verschiedene Längen. Für die Verarbeitung eines Vektors bzw. einer Matrix in der SOM ist es allerdings Vorraussetzung, dass die Länge bzw. Dimensionalität des Eingaberaumes fest vorgegeben ist und nicht während des Trainingsprozesses variiert. Um dies zu erreichen ist ein Wörterbuch aus allen relevanten Synonymgruppen erstellt worden. Im folgenden gelten also folgende Vorbedingungen. Zur Herleitung eines Wörterbuchs bestehend nur aus relevanten Synonymgruppen siehe das Kapitel Thesaurus.

1. Nachrichten liegen als Synonymgruppenvektor vor
2. Es existiert ein Wörterbuch nur aus relevanten Synonymgruppen

Eine Nachricht kann durch eine Matrix NMx (Nachrichtenmatrix) repräsentiert werden:

$$NMx = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix} \tag{5.1}$$

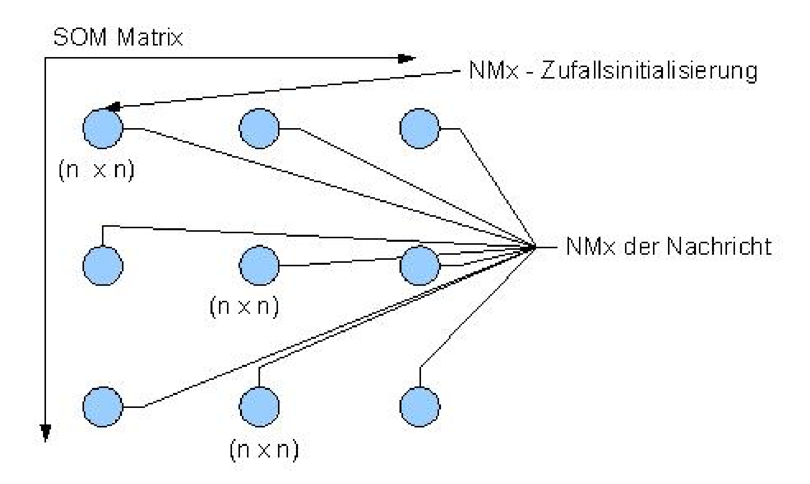


Abbildung 5.9: Self organizing map - Gewichtsmatrizen

mit n =Größe des Wörterbuchs und $w_{ij} = (P(i|j), D(i,j)), 1 \leq i, j \leq n$

- $P(i|j)$ ist die Wahrscheinlichkeit, das Synonym i vor Synonym j in einem Satz vorkommt
- $D(i,j)$ ist der durchschnittliche Wortabstand zwischen Synonym i und Synonym j

Die SOM besteht nun aus einer Gitteranordnung (Abbildung 5.9) von Knoten (Neuronen) die nun nicht mit einem Gewichtsvektor mit dem Trainingsbeispiel verbunden sind, sondern mit einer Synonymgruppenmatrix. Der Trainingsalgorithmus ist dann wie folgt:

1. Initialisiere für alle Neuronen die Synonymgruppenmatrix mit zufälligen Werten. Setze dabei die Distanz aller Elemente der Hauptdiagonalen auf 0 und verwende nur zufällige Distanzwerte aus dem Intervall [0.. Länge der bisher längsten Nachricht] und für die Wahrscheinlichkeitswerte $P(i|j) \in$ [0..1].
2. Überführe eine zufällig Nachricht (hier den Synonymgruppenvektor) in eine Synonymgruppenmatrix
3. Für jedes Neuron der SOM: für jedes Element der Synonymgruppenmatrix des Neurons berechne jeweils die neue Wahrscheinlichkeit und die neue Distanz nach folgenden Formeln
4. Hole nächste Nachricht und gehe zu Schritt 2

Das Ergebnis ist eine SOM mit automatisch adaptierten Synonymgruppenmatrizen. Die Klassifikationsphase besteht aus folgenden Schritten:

1. Überführe eine zufällige Nachricht (hier den Synonymgruppenvektor) in eine Synonymgruppenmatrix

2. Für jedes Neuron der SOM berechne den Abstand S mit:
   $S = \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \frac{P_{Neuron}(i|j)}{D_{Neuron}(i,j)} - \frac{P_{Nachricht}(i|j)}{D_{Nachricht}(i,j)} \right|$ für $i <> j$. Je kleiner S ist, desto ähnlicher sind die Matrizen. Für die grafische Darstellung trage die Zahl S in das Neuron auf der SOM Karte ein.

3. Der Benutzer sieht in welcher Region ein Neuron aktiviert ist.

Das Resultat ist, dass ähnliche Nachrichten auch Neuronen aktivieren, die nahe bei einanderliegen (im Gitter). Hinter einem Neuron verbirgt sich dabei eine Synonymgruppenmatrix. Falls ein Neuron oft Zentrum von präsentierten Nachrichten ist, bedeutet dies, dass die Synonymgruppenmatrix dieses Neurons aussagekräftig ist. Mit der Kenntnis des Anwenders, was die präsentierten Nachrichten bedeuten (Wertsteigerung, Verlust, Verkauf, Insolvenz) ist es so möglich, Wahrscheinlichkeitsverteilungen von Synonymgruppen des Wörterbuchs eindeutig einer semantischen Aussage zuzuordnen.

Anwendung dieser Methode speziell für unser System

Die SOM wurde als externes Tool, also als eigene Java Anwendung entwickelt, da im FIPs Ablauf später nur das Ergebnis (aussagekräftige Nachrichtenmatrizen) verwendet wird. Die Einbettung der SOM in das FIPs Konzept ist wie folgt:

1. Zu einem bestimmten Zeitpunkt t, für alle bisher gegebenen Nachrichten ein Wörterbuch erstellen, Nachrichten in NMx kodieren und die SOM mit den Nachrichten trainieren

2. Manuell Gruppen aus der SOM identifizieren, eine aussagekräftige Gewichtsmatrix der Gruppe als dessen Repräsentant wählen und mit einer Bewertung versehen

3. Für jede neue Nachricht nach dem Zeitpunkt t diese mit allen Gruppenrepräsentanten vergleichen und mit der Bewertung versehen, zu dessen Gruppe die Nachricht am ähnlichsten ist.

Das Programm (Abbildung 5.10) besteht aus einer graphischen Oberfläche, die in drei Regionen unterteilt ist:

1. Links ist die Gitterstruktur der Neuronen dargestellt. Jedes Neuron ist als Kreis repräsentiert. Der Abstand S ist als Zahlenwert in das jeweilige Neuron eingetragen

2. Rechts ist eine Liste aller Nachrichten (nummeriert) angezeigt.

3. In der oberen Leiste kann die Nachrichtennummer in das Eingabefeld eingegeben werden. Der Button "Los" zeigt die Aktivierung der Neuronen bezüglich der Nachricht mit der im Eingabefeld eingegebenen Nachricht

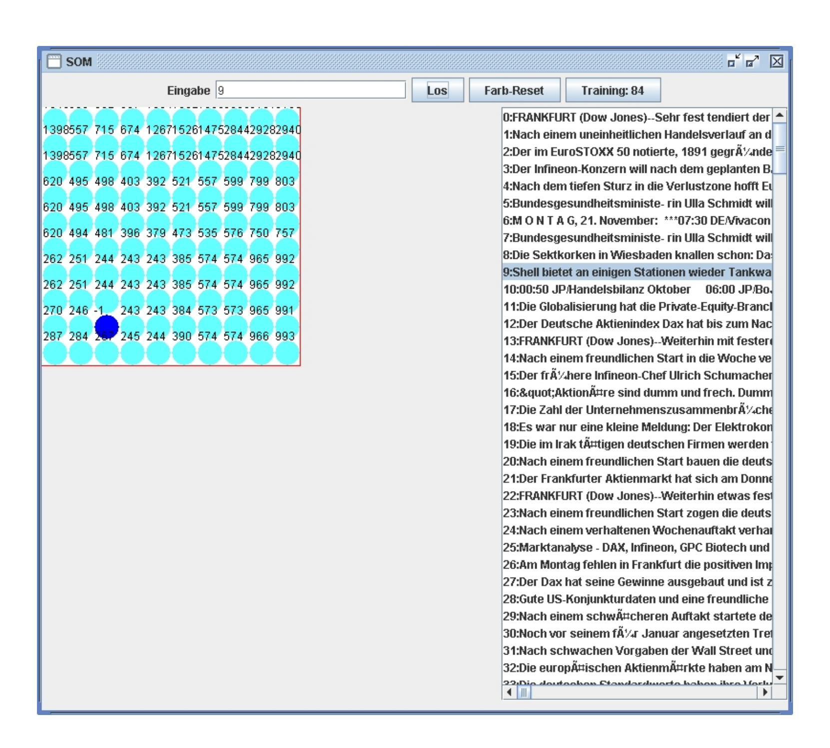


Abbildung 5.10: Self organizing map - Programm

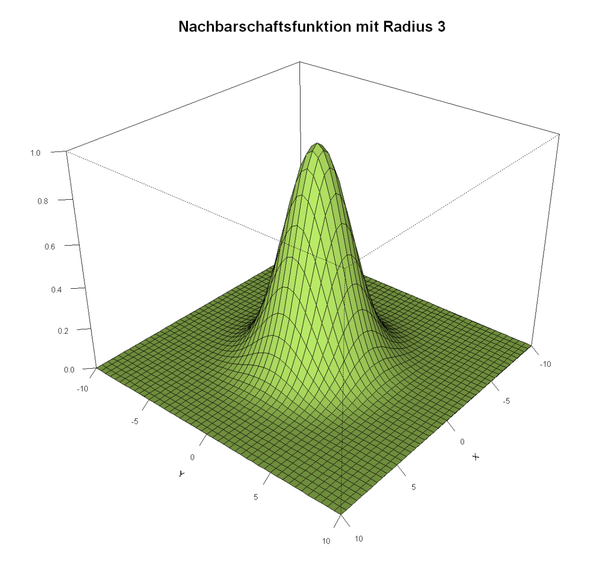


Abbildung 5.11: Self organizing map - Nachbarschaftseinfluss

an. Das Gewinnerneuron für eine zu klassifizierene Nachricht ist blau markiert und erhält als Abstandskennzeichen eine "-1". Der Button "Train" trainiert die SOM mit zwanzig zufälligen Nachrichten aus der Nachrichtenliste.

Das Programm kann wie folgt parametrisiert werden:

- Neuronenanzahl in X und Y Richtung
- Verwendetes Wörterbuch
- Lernrate
- Nachbarschaftseinflussfunktion

Die Nachbarschaftsfunktion ist als 2-dimensionale Normalverteilung realisiert mit Radius r (siehe auch Abbildung 5.11). Dabei wird an den Rändern der Gitterstruktur der Nachbarschaftseinfluss abgeschnitten, Randneuronen haben also keinen Einfluß auf die gegenüberliegenden Neuronen.

$$z = e^{\left(\left(\frac{x}{r}\right)^2 - \left(\frac{y}{r}\right)^2\right)} \tag{5.2}$$

Die Angabe der Lernrate ist parametrisiert und enthält eine Liste von Basistupeln (Anzahl Schritte, Einfluß Lernrate). Aus diesen Punkten wird mittels Newton-Basis eine Polynominterpolation vorgenommen, so dass eine (fast) monoton fallende Lernrate erzeugt wird. Das Polynom für folgende Kennzahlen ist in Abbildung 5.12 dargestellt.



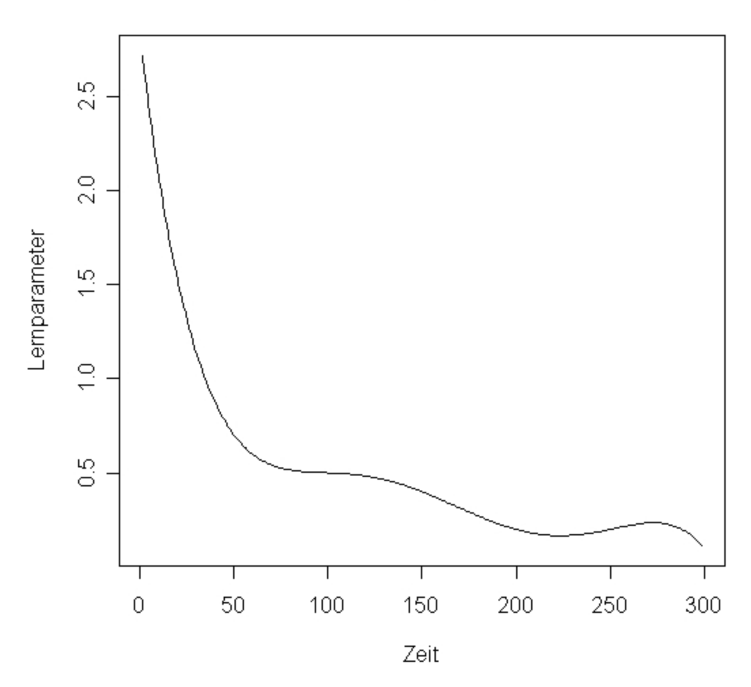


Abbildung 5.12: Self organizing map - Lernrate

Anzahl Schritte	Einfluß $p \in [0..1]$
0	0.7
1	0.7
100	0.5
200	0.4
250	0.2
300	0.1

Anwendbarkeit/Fazit

Ein Vergleich zwischen den Abbildungen 5.13 und 5.14 zeigt, wie und wo die Nachricht mit der Nummer 9 nach 21 Trainingsschritten und nach 84 Trainingsschritten klassifiziert wird. Dabei ist zu erkennen, wie die Nachbarschaftsneuronen gleichmäßig immer unähnlicher zum Gewinnerneuron werden (der Abstand wird größer). Das bedeutet, das eine Gruppierung stattfindet.

Es konnten aber aus folgenden Gründen keine eindeutigen und aussagekräftigen Gruppen gefunden werden. Aussagekräftig und eindeutig bedeuten

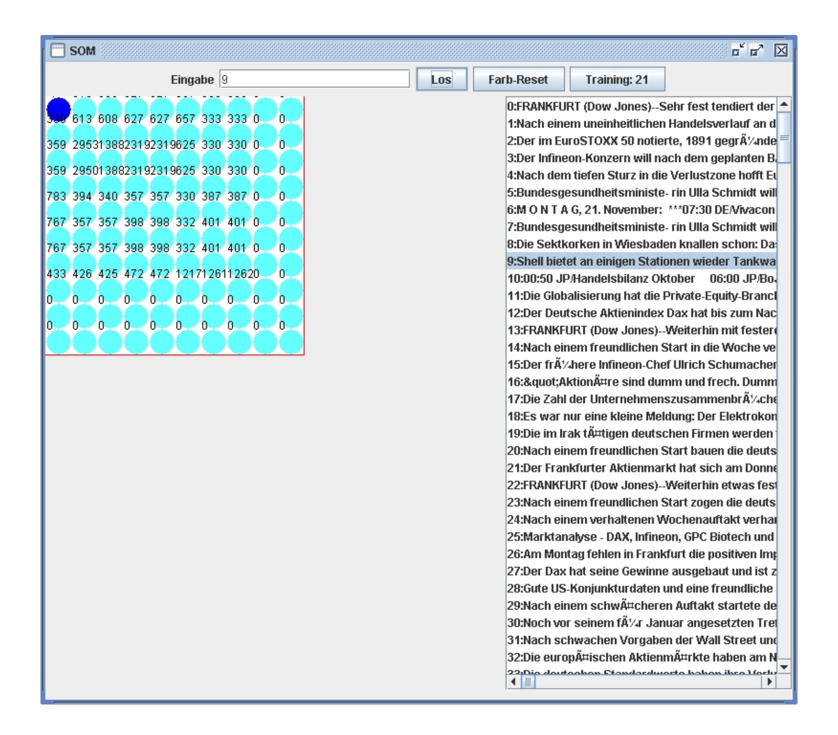


Abbildung 5.13: Self organizing map - nach 21 Trainingsschritten

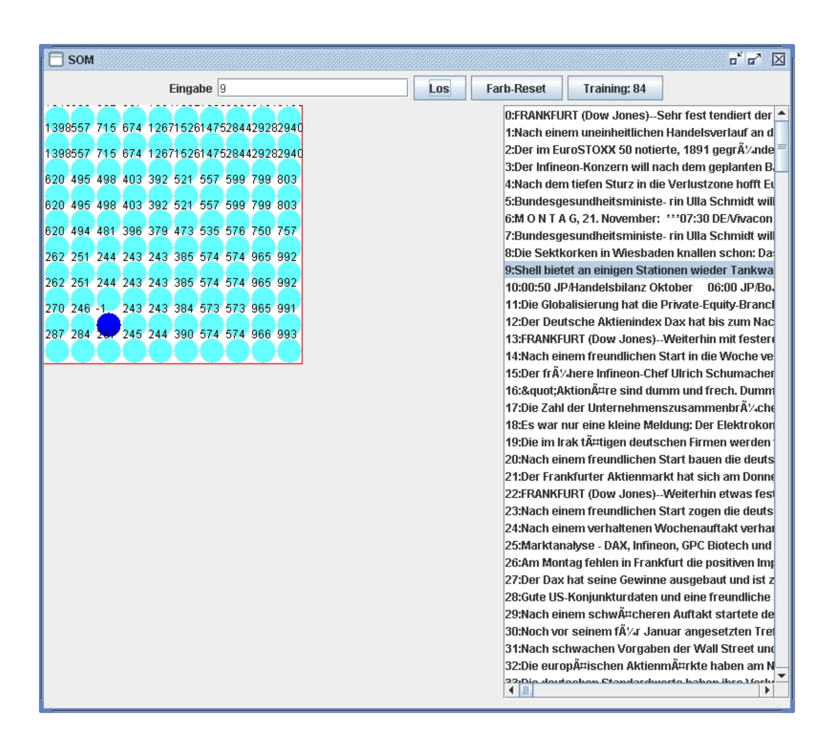


Abbildung 5.14: Self organizing map - nach 84 Trainingsschritten

hier, dass eine Gruppe möglichst nur Nachrichten enthalten sollte, die eine bestimmte wirtschaftliche Situation oder ein finanzielles Konzept wiederspiegelt, welches dann bewertet werden kann. Die Hauptgründe warum dies nicht funktioniert, liegen nicht daran, wie die SOM trainiert ist, sondern wie die Nachrichten repräsentiert sind:

1. Die Nachrichten enthalten zu viele Informationen. Es wäre besser die Nachrichten in mehrere kleinere Informationspakete aufzuteilen.

2. Es werden selten konkrekte wirtschaftliche Situationen angesprochen. Vielmehr werden diese meisst umschrieben, so dass der Kern der Nachricht nur durch semantische Analyse zu extrahieren ist.

3. Eindeutige Informationen in Nachrichten können im Zusammenhang mit anderen Nachrichten (im größeren Kontext gesehen) einen komplett anderen Sinn haben.

4. Häufig spielen Zahlen und Vergleiche von Werten und Kursen in Nachrichten eine entscheidende Rolle, um die Nachricht (manuell) zu analysieren. Alle Zahlenwerte werden bisher allerdings komplett aus dem Synonymgruppenvektor entfernt.

5. Die Laufzeit und Ressourcennutzung ist für das Programm nicht effektiv, da für jedes Neuron eine Matrix der Größe des Wörterbuchs zum Quadrat erstellt und trainiert werden muss.

5.5.7 ML-FF-Netze

Dieser Abschnitt beschreibt die Möglichkeiten der Klassifikation von Finanznachrichten durch mehrschichtige vorwärtsgerichtete künstliche Neuronale Netze. Der Kurzbeschreibung folgt das grobe Konzept, die Implementierung und schließlich die Ergebnisse der verschiedenen Testläufe.

Kurzbeschreibung

Approximation der Veränderung des Kursverhältnisses einer Aktie zum DAX bei gegebenen Finanznachrichten durch das überwachte Lernverfahren eines Multilayer Feedforward Netzes.

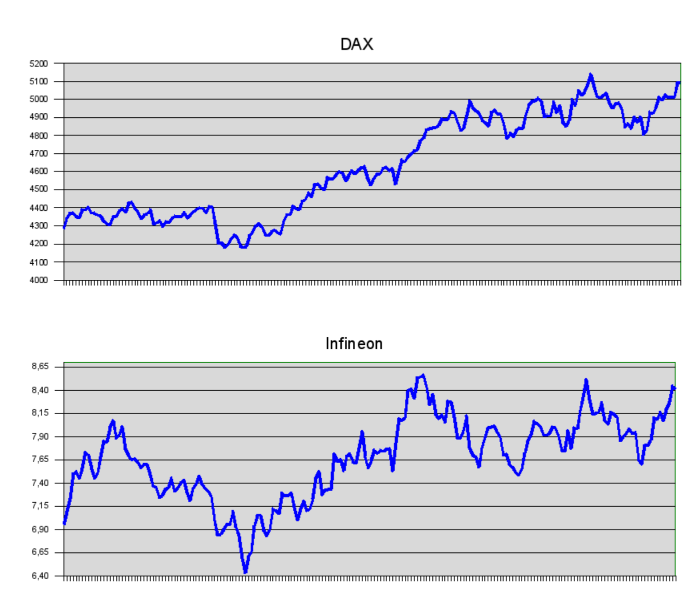
Konzept

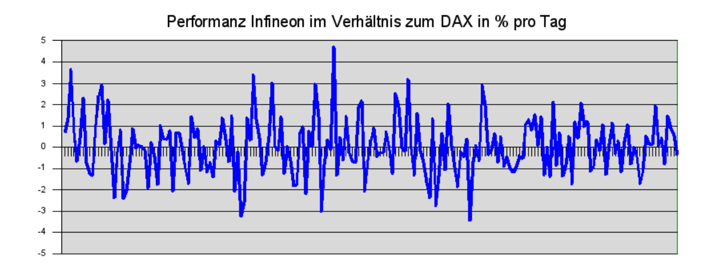
Im Prinzip suchen wir eine Möglichkeit eine Aktie in die Klassen Markt-, Out-, oder Underperformer einzuteilen. Das einzige objektive Kriterium, zu welcher dieser Klassen die Aktie letztendlich gehört, ist der Kursverlauf im Vergleich zum Marktdurchschnitt. Eine Outperformanz beispielsweise kann also nur stattgefunden haben, wenn sich der Kurs im Verhältnis zum Marktdurchschnitt besser entwickelt hat.

Gehen wir jetzt erst einmal davon aus, dass wir als einzige Informationsquelle die Finanznachrichten nutzen möchten. Das heißt wir müssen eine Verbindung zwischen einer Nachricht, die zu einer bestimmten Zeit herausgekommen ist, und der darauf folgenden Marktperformanz (nachdem diese Nachricht erschien) herstellen.

Bekannt ist, dass ein Multilayer Feedforward Netz jede unbekannte Funktion beliebig genau approximieren kann, wenn es ausreichend Neuronen gibt. Weiter reichen dafür zwei Schichten plus eine Eingabeschicht aus, da Netze mit zusätzlichen (versteckten) Schichten auf diesen Grundtyp mit einer Eingabe-, einer versteckten und einer Ausgabeschicht zurückgeführt werden können. Die Idee ist also ein derartiges Netz für die Klassifikation der Finanznachrichten zu verwenden.

Um die Marktperformanz zu bekommen reicht der Aktienkursverlauf alleine nicht aus, der Zusammenhang zum Gesamtmarkt muss hergestellt werden. Man betrachtet also das Verhältnis von Aktienkurs zum Gesamtmarkt (oder zu einem aussagestarken Marktindex wie dem DAX). Die Auswirkung einer Nachricht, und damit die zu approximierende Funktion, entspricht jetzt der Ableitung dieses Verhältnisses zum Erscheinungszeitpunkt der Nachricht, und wird im folgenden auch mit Performanz im Vergleich zum Gesamtmarkt bezeichnet. Die folgenden Abbildungen stellen den Zusammenhang am Beispiel der Aktie von Infineon im Vergleich zum DAX in der Zeit von Februar bis November 2005 dar.





Zum Training benutzen wir die in einen Vektor kodierten Nachrichten als Eingabe und die zu Klassen quantisierte Performanz im Vergleich zum Gesamtmarkt als Sollausgabe. Damit sollte das Netz in der Lage sein, Ähnlichkeiten in den Nachrichten, sowohl syntaktischer als auch semantischer Art, die zu einer gleichen Klassifizierung führen, zu lernen. Die Details der Kodierung und Quantisierung bedarf einer genaueren Untersuchung, die später folgt.

Die Wahl der Größe des Netzes - also die Anzahl der Neuronen der Eingabe-, der Ausgabe- und der versteckten Schicht - wird auch genauer untersucht. Die Neuronenanzahl der Eingabeschicht entspricht immer der Größe des Eingabevektors, die Neuronenanzahl der Ausgabeschicht sollte der Anzahl der Klassen entsprechen, so dass jedes dieser Neuronen Repräsentant einer Klasse ist. Dann wird einer neuen Nachricht mittels winner-takes-all Strategie die Klasse des am stärksten aktivierte Ausgabeneurons zugewiesen. Denkbar wäre aber auch die Klasseneinteilung "fuzzy" abzulesen, d.h. die Zugehörigkeit der Nachricht zu den einzelnen Klassen entspricht der normierten Ausgabe der jeweiligen Neuronen. Für die Wahl der Neuronenanzahl der versteckten Schicht bleibt zunächst nur Guess & Verify.

Weitere Punkte, die berücksichtigt werden müssen:

- Auswahl der Trainingsnachrichten
- Vorverarbeitung der Nachrichten durch Lemmatisierung und Ersetzung von Unternehmensnamen durch Platzhalter
- Geeignete Repräsentantion der Nachrichten
- Wieviele Trainingsnachrichten werden mindestens benötigt?

Kodierung der Finanznachrichten als Vektor

Da unser Neuronales Netz mit gewichteten Verbindungen zwischen Neuronen arbeitet, muss eine geeignete Repräsentation einer Nachricht durch einen numerischen Vektor oder eine Matrix gefunden werden.

- **naiver Ansatz**
  Ein einfacher Ansatz ist ein Wörterbuch der Länge M zu verwenden und eine Nachricht durch einen einfachen 0, 1 Vektor der Länge M anzugeben.

Jeder Stelle im Vektor ist ein Wort zugewiesen, wobei eine 1 bedeutet, dass dieses Wort im Text vorkommt und eine 0, dass es nicht vorkommt.

Probleme:
Da Vergleiche von Wörtern rein zeichenbasiert ablaufen, werden z.B. *steigt* und *steigen* als verschiedene Wörter erkannt. Das ist aus zwei Gründen nicht sinnvoll: Einmal wird das Wörterbuch dadurch unverhältnismäßig groß, zum anderen sollten sinngleiche Sätze, wie *Die Gewinne steigen weiter* und *Der Gewinn steigt weiter* zu dem gleichen Vektor führen.
Ein weiteres Problem ist, dass den Wörtern, die nur eine geringe oder gar keine Relevanz für die Aussage des Textes haben, bei der Eingabe zunächst das gleiche Gewicht zukommt, wie den wirklich wichtigen Wörtern des Textes. Es sollte, auch zur Vermeidung eines großen Wörterbuchs, eine Vorauswahl relevanter Wörter geben, was aber eher in den Problembereich fällt, ein gutes Wörterbuch zu erstellen.

- **Synonymgruppen-Vektor Ansatz**
  Dieser Ansatz baut auf dem naiven Ansatz auf, benötigt jedoch eine Reihe an vorverarbeitenden Schritten.
  Der Eingabevektor für das Netz ist wieder ein 0, 1 Vektor der Länge M, jedoch diesmal über einem Wörterbuch aus Synonymgruppen. So ein Wörterbuch enthält M Zahlenwerte, wobei jede dieser Zahlen für eine Synonymgruppe[5] steht. Zwei oder mehr Wörter sind Synonyme, wenn sie in einem bestimmten Kontext die gleiche Bedeutung haben. Die Synonymgruppe 319 besteht z.B. aus „Ausbeute, Einnahmen, Erlös, Ertrag, Gewinn, Profit, Rendite, Überschuss".
  Um zu dieser Darstellung zu kommen, müssen die Finanznachrichten einige preprocessing-Schritte durchlaufen. Für eine genauere Beschreibung siehe Allgemeine Methoden.

  1. **Stemming / Lemmatisierung:**
     Wörter werden auf ihre Stammform zurückgeführt.
  2. **Stoppwort Bereinigung:**
     Text wird um sogenannte Stoppwörter (Wörter die sehr häufig in Texten vorkommen) bereinigt.
  3. **Synonymgruppen finden:**
     Zu jedem verbleibenden Wort des Textes wird die Synoymgruppe herausgefunden.

  Probleme:
  Dieses Vorgehen nimmt keinerlei Bezug auf die Position der Wörter im Text und auf den Kontext. Die Rekonstruktion der Nachricht, die dann aus einer Menge von Synonymgruppen besteht, deutet nur noch teilweise auf den Inhalt hin, es kann aber kein Bezug zwischen Prädikaten und Objekten

[5]Synonymgruppe entspricht meaning_id aus dem Open Thesaurus, für weiter Informationen siehe www.openthesaurus.de

hergestellt werden. Trotzdem besteht die Hoffnung, dass Kernaussagen erkannt werden.

- **Berücksichtigung des Kontextes**
  Eine Repräsentation, die den Kontext berücksichtigt, ist die Folgende:
  - "sinnvolles" Wörterbuch von Synonymgruppen der Größe M, aber nicht zu groß (max. 100 Gruppen)
  - 2-dimensionales Array der Größe MxM
    $$X = \begin{bmatrix} x_{1,1} & \cdots & x_{M,1} \\ \vdots & \ddots & \vdots \\ x_{1,M} & \cdots & x_{M,M} \end{bmatrix} \text{ mit } x_{i,j} = \begin{cases} 1 & \text{,falls Wort j im Text direkt hinter Wort i} \\ 0 & \text{,sonst} \end{cases}$$
  - Wörterbuch der Form:

4254	Zeile 1 (steht für Synonymgruppe: nicht)
6235	Zeile 2 (steht für Synonymgruppe: Vorjahr)
12345	Zeile 3 (steht für Synonymgruppe: Quartalszahlen)
6235	Zeile 4 (steht für Synonymgruppe: Gewinn)
3577	Zeile 5 (steht für Synonymgruppe: Verlust)
7836	Zeile 6 (steht für Synonymgruppe: steigen)
8355	Zeile 7 (steht für Synonymgruppe: fallen)

  - Die Größe des Wörterbuches entspricht der Zeilenanzahl. In jeder Zeile steht die Nummer einer Synonymgruppe.
    Beispiel:
    "Gewinn" sei in der Synonymgruppe in Zeile 4, "fallen" in Zeile 7.
    "Der Gewinn fällt" hat dann in der Matrix die Stelle $x_{4,7} = 1$.

  Vorteil

  - Erhält die Syntax teilweise, da Nachbarschaften von Wörter gespeichert werden. Falls kein Wort im Text doppelt vorkommt ist eine Rekonstruktion des Textes (aus den Synonymgruppen im Wörterbuch) möglich.

  Nachteile

  - Bei mehrfachem Vorkommen der gleichen Synonymgruppe ist keine eindeutige Rekonstruktion des Textes möglich.
  - Große Eingabematrix und große Anzahl an Gewichten:
    bei $M = 100$ schon 10.000 double Werte pro Matrix
    bei $10.000 + 200 + 7$ Neuronen in den Schichten:
    Anzahl Gewichte: $10.000 \cdot 200 + 200 \cdot 7 = 2.001.400$ double Werte
  - Lösungsansätze:
    * Entwurf eines sehr strengen Wörterbuchs oder
    * mit der langen Rechenzeit leben.

Quantisierung der Performanz im Vergleich zum Gesamtmarkt

Da unser Neuronales Netz als Klassifizierer arbeiten soll, müssen vorher Klassen festgelegt werden, in die die Finanznachrichten eingeordnet werden können. Wie schon beschrieben dient als Grundlage der Klasseneinteilung die Performanz der Aktie im Vergleich zum Gesamtmarkt (PAzuG), in einem gewissen Zeitraum nach Erscheinen der Nachricht.

- **Wahl eines geeigneten Wirkungszeitraums einer Nachricht**
  Geht man davon aus, dass Nachrichten ausschließlich zum Erscheinungszeitpunkt wirken, so wäre die Ableitung der PAzuG ein geeignetes Maß, da sie genau die Stärke der Änderung der PAzuG zeigt.
  Ein weiterer Ansatz könnte die durschnittliche Änderung der PAzuG bis zum Erscheinen der nächsten Nachricht zum jeweiligen Unternehmen sein. Der Praxis angemessener ist, die Änderung der PAzuG über eine Zeitraum zu untersuchen, der als realistischer Wirkungszeitraum der Nachricht angenommen werden kann. Nachrichten werden zuerst stärker beachtet, die Wirkung schwächt sich aber mit der Zeit ab. Was dabei "realistisch" bedeutet, ist im Zweifelsfalle durch Testen verschiedener Parameter herauszufinden.

- **Änderung der PAzuG über Wirkungszeitraum**
  Voraussetzungen

  - Es werden erst einmal nur Aktien aus dem DAX betrachtet und der DAX selber als geeigneter Repräsentant der Gesamtmarktentwicklung gesehen.
  - Sei t der Zeitpunkt des Eintreffens einer Nachricht und t_0 vor dem Eintreffen und t_1,t_2,t_3 mit $t < t_1 < t_2 < t_3$ Zeitpunkte nach dem Eintreffen.
  - Benutze 7 Klassen: -3 , -2 , -1 , 0 , 1 , 2 , 3 (von starker Underperformance über Markt- bis starker Outperformanz)

  Einteilung in Klassen

  - Verhältnis v_0 von Aktienkurs zum DAX zum Zeitpunkt t_0 wird berechnet: Damit ist das Verhältnis vor dem Einfuß der Nachricht bekannt.
  - Analoge Berechnung von v_1, v_2, v_3 für die Zeitpunkte t_1, t_2, t_3.
  - Setze Bewertung zuerst auf Null: Bewertung = 0;
  - Jetzt werden die Verhältnisse mit v_0 verglichen.
    * Ist $v_0 \leq v_1 \cdot (1 - tol)$, so ist der Kurs der Aktie stärker gestiegen als der Gesamtmarkt, die Aktie hat den Gesamtmarkt Outperformed. Die Bewertung wird um eins erhöht. Dabei ist tol der Toleranzwert, z.B.: $tol = 0,002$, dann gilt die Aktie nur als Outperformer, wenn sie sich um mindesten $0,2\%$ besser als der Gesamtmarkt entwickelt.

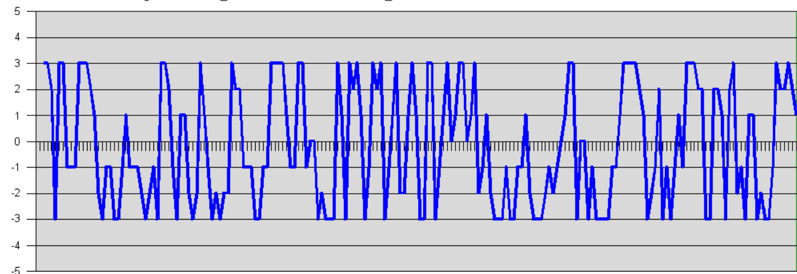
    * Ist $v_0 \geq v_1 \cdot (1 + tol)$, so hat die Aktie underperformed und die Bewertung wird um eins verringert.

    * Ist $v_0 > v_1 \cdot (1 - tol)$ und $v_0 < v_1 \cdot (1 + tol)$, so war die Veränderung nur gering und die Performanz entspricht in etwa der des DAX. Die Bewertung bleibt wie sie ist.

- Analog werden v_2 und v_3 mit v_0 verglichen.

Am Ende ergibt sich eine Bewertung, die auch dem intuitiven Verständnis entspricht. Hat die Aktie den DAX dreimal outperformed, war die Nachricht stark positiv, im umgekehrten Falle eben stark negativ. Eine geeignete Wahl für t_1, t_2 und t_3 könnten $t_1 = t + 1$ Handelstag, $t_2 = t + 2$ Handelstage, $t_3 = t + 5$ Handelstage sein.

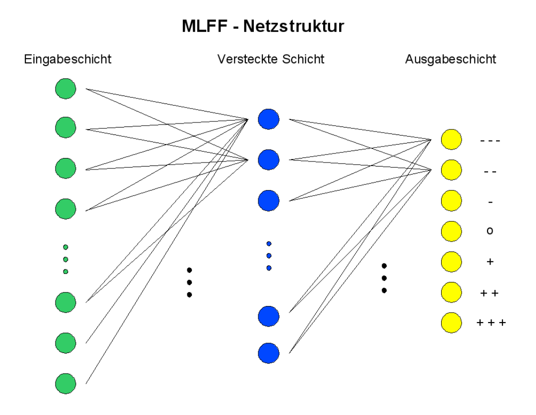
Die Abbildung zeigt die berechnete Klasseneinteilung einer Nachricht zu Infineon für den jeweiligen Handelstag.



Damit stehen die Klassen, die den Nachrichten im Rückblick zugeordnet werden können, fest. Nach beendetem Training werden auch neue Nachrichten klassifiziert. Der Erfolg des Netzes ist qualitativ meßbar, indem man die Prognose vom Netz später mit der realen Klasse vergleicht.

Implementierung

Es wurde ein feedforward Netz mit einer Eingabe-, einer versteckten und einer Ausgabeschicht implementiert. Dieses Netz soll Finanznachrichten klassifizieren, die in der oben beschriebenen Synonymgruppen-Vektor Form kodiert sind. Die Struktur ist in folgender Abbildung erkennbar:



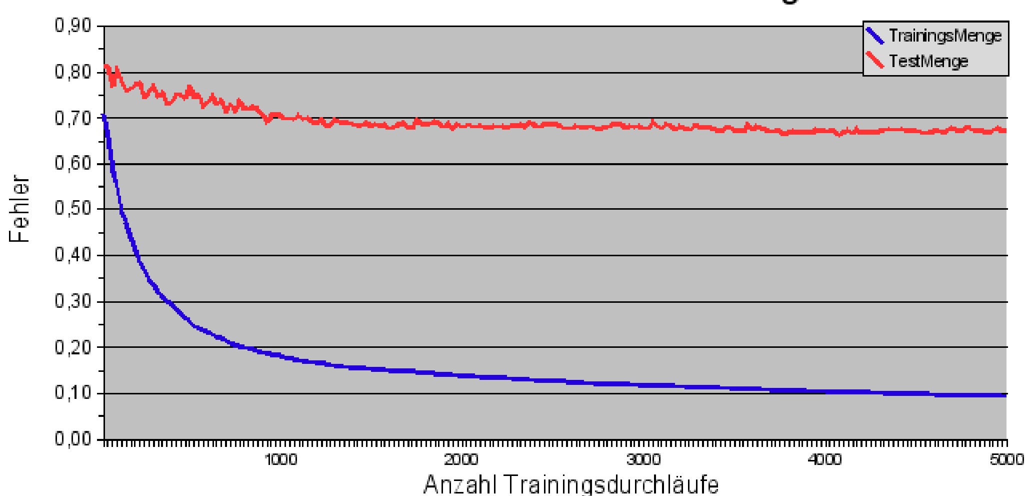
- Anzahl Neuronen in der Eingabeschicht = Wörterbuchgröße = 1306
- Anzahl Neuronen in der versteckten Schicht = 100
- Anzahl Neuronen in der Ausgabeschicht = Anzahl verschiedener Klassen = 7
- Aktivierungsfunktion der Neuronen der versteckten und der Ausgabeschicht ist die sigmoide (s-förmige) logistische Funktion: $sgd(x) = \frac{1}{1+e^{-x}}$. Ihre Ableitung $\frac{sgd}{dx}(x) = \frac{e^{-x}}{(1+e^{-x})^2}$ kann durch die Funktion selbst wieder ausgedrückt werden $\frac{sgd}{dx}(x) = sgd(x) \cdot (1 - sgd(x))$, was die Berechnung bei der Anpassung der Gewichte vereinfacht.
- Ausgabefunktion ist Identität
- Trainingsart: Backpropagation im Batch Modus mit je 10 Trainingsbeispielen pro Etappe
- Die Lernrate ist monoton fallend zur Anzahl der Trainingsbsp (anzBsp): $0,6 \cdot 0,988^{(\text{anzBsp})} + 0,1$
- Feature: Gewichtsanpassung mit Momentum Term für schnellere Konvergenz mit $\alpha = 0,5$

Test

Zum Testen werden 353 bewertete Beispielnachrichten zu DAX Unternehmen benutzt, die nach den Verfahren aus Abschnitt 8.5 Testdaten erstellt wurden. Von diesen Nachrichten werden 326 zum Training und 27 zum Test des Netzes verwendet. Das dazu erstellte Wörterbuch enthält 1306 Synonymklassen.

Das Netz wurde in 5000 Trainingsetappen mit den Trainingsdaten trainiert. Dafür brauchte ein Notebook mit 1,7 Ghz Centrino Prozessor und 512 MB RAM circa eine Stunde. Nach je 25 Etappen wurde der mittlere quadratische

Fehler zwischen Soll- und Ist-Ausgabe über alle Trainings- und über alle Testdaten berechnet. Der Verlauf dieser beiden Fehler ist in folgender Abbildung dargestellt.



Wie man sieht, lernt das Netz die Trainingsdaten relativ schnell. Ab 3000 Etappen etwa ist der Fehler in einem annehmbaren Bereich. Für die Testdaten sinkt der Fehler aber im gleichem Zeitraum kaum. Das liegt daran, dass die Testdaten zu verschieden zu den Trainingsdaten sind. Die neuen Eingabemuster passen nicht zu den gelernten.

Dass gelernte Nachrichten nur eine geringe Aussage für neue Nachrichten haben, sieht man auch an folgendem Test:

Es wurden für alle Trainings- und Testbeispiele auf dem fertig trainierten Netz das jeweilige Gewinnerneuron und dessen Ausgabewert berechnet. War die Ausgabe ≥ 0.6, so wurde die Klassifikation als *sicher* angenommen, und berechnet, ob das Netz richtig oder falsch lag. Lag die Ausgabe unter 0.6 wurde die Nachricht als *unsicher* klassifiziert und berechnet, ob das Netz trotzdem richtig oder falsch gelegen hätte. Das Ergebnis zeigt folgende Tabelle:

	Train	Test	gesamt
Anzahl Tests	326	27	353
richtig klassifiziert:	299 (91,7%)	8 (29,6%)	307 (87,0%)
falsch klassifiziert:	1 (0,3%)	11 (40,7%)	12 (3,4%)
als unsicher klassifiziert:	26 (8,0%)	8 (29,6%)	34 (9,6%)
davon wären richtig:	7 (26,9%)	5 (62,5%)	12 (35,3%)
davon wären falsch:	19 (73,1%)	3 (37,5%)	22 (64,7%)

Wie zu erwarten, klassifizierte das Netz fast alle Trainingsnachrichten richtig, von den *unsicheren* wären aber viele falsch gewesen, sodass ein Grenzwert, ab dem eine Nachricht als unsicher klassifiziert gilt, durchaus sinnvoll ist.
Von den Testnachrichten hingegen wurden nur knapp ein Drittel (29, 6%) richtig klassifiziert. Immerhin war das Netz bei diesem Test besser als der Zufall mit einem erwarteten Wert von 1/7, also ca. 14, 3% an richtigen Zuordnungen.

Hier sieht man wieder wie komplex der Themenbereich der Klassifikation von Finanznachrichten ist. Um die gesamte Methode der Klassifikation mit Neuronalen Netzen weiter zu testen sind also viele weitere Trainingsnachrichten nötig, um so annähernd das Spektrum aller möglichen Finanznachrichten abzudecken. Nur so werden sinnvolle Aussagen erst möglich.

Anwendbarkeit/Fazit

Die tatsächliche Anwendbarkeit der Methode ist noch nicht endgültig geklärt. Vor allem mangelt es an geeigneten vorverarbeiteten Trainingsnachrichten, um genauere Untersuchungen zuzulassen. Es sprechen jedoch eine Reihe von Begründungen gegen einen wirklich überragenden Erfolg dieses Ansatzes.

- Die Komplexität einer Finanznachricht scheint nicht einfach in einen linearen Vektor von Synonymgruppen komprimierbar zu sein. Insbesondere wird kein Bezug auf die Semantik genommen.
- Der Ansatz basiert darauf, dass aus vorhandenen (trainierten) Nachrichten Aussagen über neue Nachrichten geschlossen werden sollen. Ein Schwachpunkt hierbei ist, dass die berechnete Aussage bei einer zu unähnlichen neuen Nachricht wertlos ist, da die (beim Training erstellten) Funktionen für die Berechnungen auf eine andere Situation abzielen. Es kann aber nicht erkannt werden, wann eine Nachricht zu unähnlich ist.
- Eine Erweiterung des Ansatzes mit einer Eingabematrix, die den Kontext der Wörter berücksichtigt (wie oben beschrieben), würde das Problem der Unähnlichkeit der Nachrichten nur vergrößern. Außerdem erreicht man durch die quadratische Größe bzgl. des Wörterbuches schnell die Performancegrenze.

- Finanznachrichten müssen meistens im Kontext des Marktes und den Erwartungen der Anleger gesehen werden. Sogar Wörter wie *Rekordgewinn* sind nicht zwingend positiv, falls dennoch die (hohen) Erwartungen der Anleger damit enttäuscht werden (besonders in schnell wachsenden Branchen gültig).
- Da die Einschätzung von Finanzprodukten im Wesentlichen auf Zahlen **und** Fakten beruht, und wir bestenfalls die Fakten erfassen können, ist der Erfolg zweifelhaft.

Kapitel 6

Entscheidungsfindung (Grundlage: techn. Analyse)

6.1 Einleitung

In der Projektgruppe haben wir uns dafür entschieden, für die Finanzprodukte eine ,,technische Analyse" zu realisieren, welche auf Fundamentalkennzahlen zu einzelnen Unternehmen basiert. Eine Analyse des Kursverlaufs einer Aktie wurde nicht vorgenommen.

In Kapitel 3 und 4 haben wir uns darüber Gedanken gemacht, wie man einen einzelnen Kunden und ein einzelnes Finanzprodukt sinnig repräsentieren kann, so dass diese Daten zur Weiterverarbeitung in unserem System geeignet sind.
Ziel ist es nun, in der Entscheidung aufgrund technischer Fundamentaldaten, dem Kunden, welcher sich durch einen spezifischen Kundenvektor charakterisieren lässt, ,,gute" Finanzprodukte zu empfehlen.

Die Güte eines Finanzproduktes für einen bestimmten Kunden ist nicht nur durch die vorausberechnete Rendite zu bestimmen. Finanzprodukte mit höherer Rendite haben zumeist auch eine geringere Sicherheit, so dass für einen Kunden somit auch das Risiko steigt, das eingesetzte Kapital zu verlieren bzw. zu minimieren.

Ziel muss es demnach sein, einem Kunden das für sein Anlageprofil passende Finanzprodukt (in unserem Projekt konzentrieren wir uns auf Aktien) zu empfehlen, und die Aktien, die wir betrachten (DAX30), in einem Ranking darzustellen.
Das Anlageprofil des Kunden entnehmen wir dem Kundenvektor *((Sicherheits,*

Verfügbarkeits, Rendite) - Wunsch des Kunden) aus Kapitel 3 und die Einschätzung des Finanzproduktes *(Sicherheit, Verfügbarkeit, Marketperformer)* entnehmen wir dem berechneten Sicherheitswert aus Kapitel 4, setzen die Verfügbarkeit auf einen festen Wert *(die Verfügbarkeit ist für alle Aktien gleich: ,,kurzfristig verfügbar")* und berechnen anschließend die Rendite anhand der Gesetzmäßigkeiten des magischen Dreiecks.

Aufgrund der Berechnungen des Kundenvektors, welcher für jedes Attribut eine Zahl zwischen 0 und 100 darstellt, macht es keinen Sinn, eine feste Grenze zu ziehen, ab welcher ein Kunde einem bestimmten Bereich zugeordnet wird. Es könnte sein, dass er nur einen Punkt neben der Grenze liegt, die zum nächsten Bereich gehören würde. Diese scharfe Einteilung würde demnach keinen Sinn ergeben, da ein Kunde in solchen überschneidenden Fällen eher mehreren Klassen zugeordnet werden sollte, und das mit unterschiedlicher Zugehörigkeit.
An diesem Punkt kommt die Fuzzy-Logik ins Spiel, welche eine unscharfe Logik darstellt und genau jenes leistet: Ein Kunde kann mehreren Sicherheits-, Verfügbarkeits- und Rendite-Klassen angehören, und das mit unterschiedlicher Zugehörigkeit.
Gleiches gilt für die Finanzprodukte: Ein Finanzprodukt kann mehreren Sicherheits- und Marketperformer-Klassen angehören, und das mit unterschiedlicher Zugehörigkeit. Die Verfügbarkeit allerdings ist für ein Finanzprodukt, zumindest für unser Projekt (wir betrachten nur Aktien), immer gleich und kann somit durch eine scharfe Grenze oder einen scharfen Wert eingestellt werden.

Die Entscheidung hat nun die Aufgabe, den Kunden und das Finanzprodukt mittels Fuzzy-Logik bestimmten Klassen zuzuordnen, welche aus Tupeln von Sicherheit, Verfügbarkeit und Rendite bestehen, und diese Klassen danach sinnig zu verschmelzen, so dass aus der *Übereinstimmung* zwischen den Klassen des Kunden und denen des Finanzproduktes ein **Rankingwert** berechnet werden kann. Ausgehend von dieser Berechnung kann einem Kunden so nun zu jedem Finanzprodukt die Übereinstimmung zu seinem Anlageprofil berechnet werden, und das ist die eigentliche ,,Entscheidung".

6.2 Fuzzy Logik Einführung

6.2.1 Motivation

Die Fuzzy Logik grenzt sich von der ,,klassischen" Logik insofern ab, als dass sie **,,unscharfes Schließen"** erlaubt und explizit mit den Mechanismen des unscharfen Schließens rechnet.

Mittels Fuzzy-Mengen können die ,,feinen" Bereiche des Kunden bzw. Finanzvektors dann gröberen Beschreibungen zugeordnet werden, und zwar mit gewissen Toleranzen, so dass nicht immer nur ein starrer Bereich vorgegeben ist, sondern darüber hinaus auch andere Bereiche in einer ermittelten Ordnung

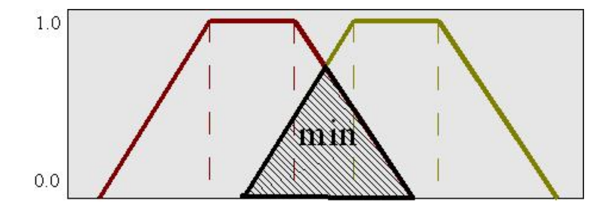


Abbildung 6.1: Fuzzy-Minimum Operation

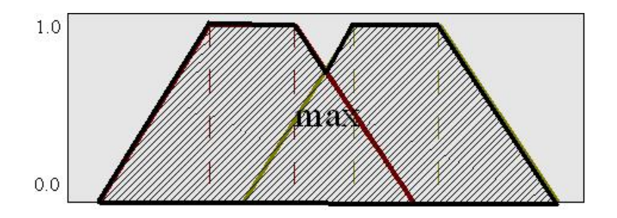


Abbildung 6.2: Fuzzy-Maximum Operation

zugeordnet werden können.

6.2.2 Operationen

Mithilfe der T- und S-Normen und der *Max bzw. Min* Operatoren fassen wir Fuzzy Mengen zusammen.

min-Operator

Für zwei Fuzzy Sets A und B ist der Minimum Operator definiert zu:

$$\mu_c(x) = \min\{\mu_A(x), \mu_B(x)\}$$

Anschaulich:
A = Sicherheit sehr gering, B = Sicherheit gering
$\Longrightarrow$ C = Sicherheit ist in A geschnitten B, also : Sicherheit ist sehr gering **und** gering (logisches „und").

max-Operator

Für zwei Fuzzy Sets A und B ist der Maximum Operator definiert zu:

$$\mu_c(x) = \max\{\mu_A(x), \mu_B(x)\}$$

Anschaulich:
A = Sicherheit sehr gering, B = Sicherheit gering
$\Longrightarrow$ C = Sicherheit ist in A vereinigt B, also : Sicherheit ist sehr gering **oder** gering (logisches „oder").

und-gamma-Operator

In diesem Unterkapitel geht es um (komplexe) Kombinationen von Rendite, Sicherheit und Verfügbarkeit (2-3 Dimensionen).

Der Sinn dabei, diese und-gamma-Funktion zu verwenden, liegt darin begründet, dass man später (bei der Kombination von Kunde und Finanzprodukt (Ranking)) feststellen möchte, in welche Kundenklasse ein Kunde fällt und mit welcher Zugehörigkeit.
Da ein Kunde aus einem Vektor mit 3 Komponenten besteht (Sicherheitswert, Verfügbarkeitswert, Renditewert), müssen diese 3 Komponenten, welche durch Fuzzy-Logik umgesetzt werden, geeignet mit einer Fuzzy-Funktion aggregiert werden, um die Zugehörigkeit eines Kunden zu einem Tupel aus (Sicherheit, Verfügbarkeit, Rendite) bestimmen zu können.
Für jedes mögliche Tupel wird dann eine Zugehörigkeit durch die und-gamma-Funktion angegeben und wir können bestimmen, welche Tupel diejenigen mit maximaler Zugehörigkeit sind. Diese sind dann dazu geeignet, den Kunden bestmöglich zu repräsentieren.

Genereller Unterschied: Die Funktionen sind nicht mehr in einer Dimension sondern in zwei Dimensionen (z.B. Sicherheits- und Verfügbarkeits-Funktionen können unterschiedliche Werte enthalten, genauer: Sicherheitswert vom Kunden ist 21 und Verfügbarkeitswert ist 54).

Die eingesetzten Funktionen $\mu_{Sicherheit=hoch}$ und $\mu_{Verfuegbarkeit=mittelfristig}$ verarbeiten nicht den gleichen Wert, denn in Sicherheit = hoch wird der Sicherheitswert und in der anderen der Verfügbarkeitswert eingetragen. Somit entsteht ein 2-dimensionales Feld an Werten (Tupeln), denen dann jeweils Fuzzy-Werte zwischen 0 und 1.0 durch die Fuzzy-Funktionen zugeordnet werden. Das geht am besten mit „Fuzzy Und gamma", da nach Betrachtung des Schaubildes der Funktion sich diese als geeignet für unsere Klassifizierung herausgestellt hat.

Durch diese Kombinationen können die drei Dimensionen des Kundenvektors (Sicherheit, Verfügbarkeit und Rendite) untereinander zusammengefasst werden, so dass Klassen entstehen, die den Kunden kategorisieren können (z.B. welchen Zugehörigkeitswert hat der Kunde zu der Klasse, bei der Sicherheit hoch, Verfügbarkeit kurzfristig und Rendite mittel ist ?). Diese Werte könnten dann absteigend sortiert werden, so dass für den Kunden mehrere Klassen mit unterschiedlicher Zugehörigkeit entstehen können, wodurch auch die zu empfehlenden Aktien abhängen. Wenn man diese Klassen noch erweitern möchte, dann könnte man z.B. die Sicherheitsklassen vorher noch durch elementare Kombinationen erweitern und diese erweiterten Fuzzy-Mengen dann in die Kombinationsfunktion einsetzen. Somit sind die Möglichkeiten (fast) unbegrenzt.

Kombinieren von 2 Dimensionen:

Fuzzy ,,Und gamma" ist definiert zu:

$$\mu(\mu_A(x), \mu_B(y)) = \gamma * \min(\mu_A(x), \mu_B(y)) + \frac{1}{2} * (1-\gamma) * (\mu_A(x) + \mu_B(y)) \text{ für } \gamma \in [0,1]$$

Als gamma Wert wird 0.5 gewählt; dieser scheint sich am besten zu eignen (nach Betrachtung der Diagramme dieser Funktion), also:

$$\mu(\mu_A(x), \mu_B(y)) = \frac{1}{2} * \min(\mu_A(x), \mu_B(y)) + \frac{1}{4} * (\mu_A(x) + \mu_B(y))$$

Beispiel:

A = Sicherheit ist mittel, B = Verfügbarkeit ist kurzfristig (x ≡ Sicherheitswert des Kunden, y ≡ Verfügbarkeitswert des Kunden).

$\mu(\mu_A(x), \mu_B(y))$ ≡ Zugehörigkeitswert des Kunden zu der Klasse, bei der Sicherheit mittel ist und Verfügbarkeit kurzfristig.

Kombinieren von 3 Dimensionen:

Da ,,Fuzzy-Und" nicht assoziativ ist, kann das Assoziativgesetz nicht eingesetzt werden und somit die Funktion für 2 Dimensionen nicht angewendet werden. Für 3 Dimensionen wird die Funktion von uns neu definiert, so dass sie nun kommutativ ist:

$$\mu(\mu_A(x), \mu_B(y), \mu_C(z)) = \gamma * \min(\mu_A(x), \mu B(y), \mu_C(z)) + 1/3 * (1-\gamma) * (\mu_A(x) + \mu_B(y) + \mu_C(z)) \gamma \in [0,1]$$

Als gamma Wert wird wiederum 0.5 gewählt, also:

$$\mu_(\mu_A(x), \mu_B(y), \mu_C(z)) = 0,5 * min(\mu_A(x), \mu_B(y), \mu_C(z)) + 1/6 * (\mu_A(x) + \mu_B(y), \mu_C(z))$$

Beispiele

Die Fuzzy-Operationen können auch in einer anderen Darstellungsart angezeigt werden: In einem Quadrat werden auf einer Skala von 0 bis 1.0 auf zwei Achsen Werte abgetragen für die Fuzzy-Sets A und B. In der Mitte wird zu jeweils einem Wert aus A und B der Funktionswert der jeweiligen Fuzzy-Funktion abgetragen. Der Funktionswert wird anhand einer Skala in Graustufen abgetragen, die rechts im jeweiligen Bild zu sehen ist.

Die Grafiken sollen verdeutlichen, wie sich die Werte der entstandenen Funktion berechnen. Besonders intensiv haben wir uns mit der Und-Gamma-Funktion beschäftigt, deren Schaubild uns als geeignet für die Kombination erschienen ist.

Die folgenden Bilder sind folgendem Buch entnommen:

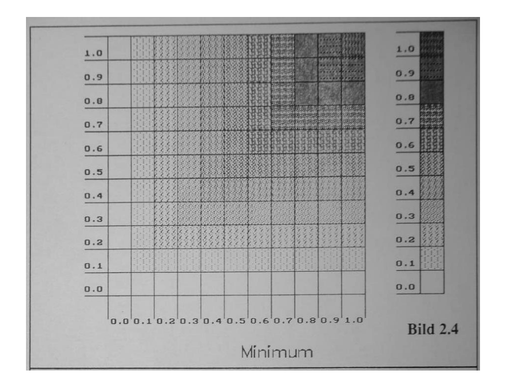


Abbildung 6.3: Fuzzy-Minimum Operator als Schaubild

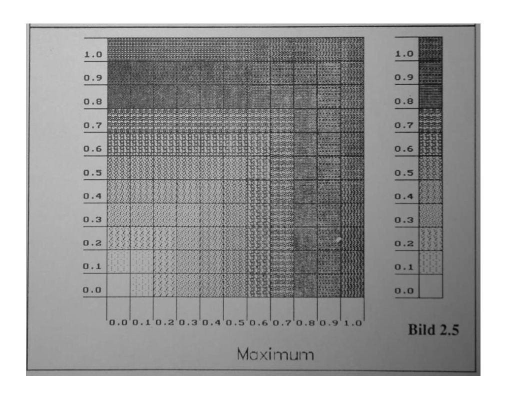


Abbildung 6.4: Fuzzy-Maximum Operator als Schaubild

Fuzzy-Logik: Grundlagen, Anwendungen, Hard-und Software - Thomas Tilli - 2., unver. Aufl. - München: Franzis, 1992

6.2.3 Fuzzymengen

Fuzzy-Mengen (oder Fuzzy-Sets) sind, wie oben beschrieben, Abbildungen, die jedem Element $x \in X$ einen Zugehörigkeitsgrad $\mu(x) \in [0, 1]$ zuordnen.
Es gibt ansonsten keine weitere Vorschrift, wie diese Fuzzy-Mengen auszusehen haben. Dennoch gibt es einige Grund-Funktionen, mit denen sich diese Zugehörigkeitsfunktion angeben lässt. Diese Funktionen seien im Folgenden kurz vorgestellt.

Dreieck

Das Dreieck besteht aus einem Mittelpunkt m, bei dem der Zugehörigkeitswert maximal wird. Zudem benötigt dieses Verfahren noch zwei Parameter, α und β, welche jeweils die Länge des ,,Anstiegs" vom Wert 0.0 zum Mittelpunkt bzw.

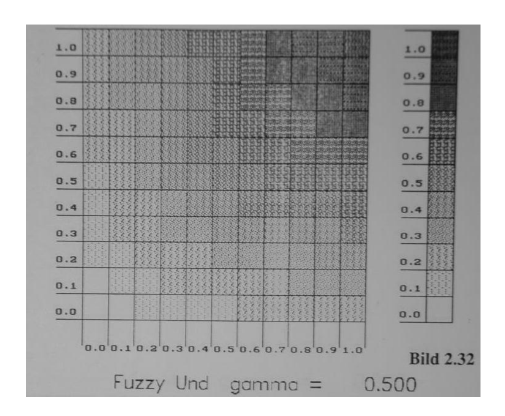


Abbildung 6.5: Fuzzy-Und-Gamma Operator als Schaubild mit $\gamma = 0.5$

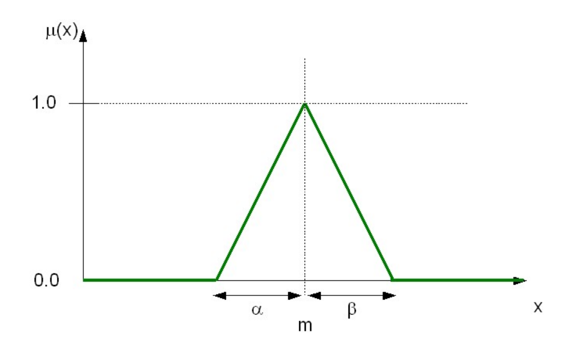


Abbildung 6.6: Fuzzy-Dreieck - Funktion

,,Abstiegs” vom Mittelpunkt auf den Wert 0.0 angeben.

Ein Beispiel dieser Funktion ist in Abb. 6.6 gegeben.

Trapez

Das Trapez besteht aus zwei ,,Mittelpunkten”, m_1 und m_2, und den bereits bekannten Parametern α und β.
Zwischen m_1 und m_2 beträgt der Wert der Zugehörigkeitsfunktion 1.0. α gibt die Länge des Anstiegs vom Wert 0.0 auf 1.0 zum Punkt m_1 an und β gibt die Länge des Abstiegs von dem Punkt m_2 (1.0) auf 0.0 an.

Ein Beispiel dieser Funktion ist in Abb. 6.7 gegeben.

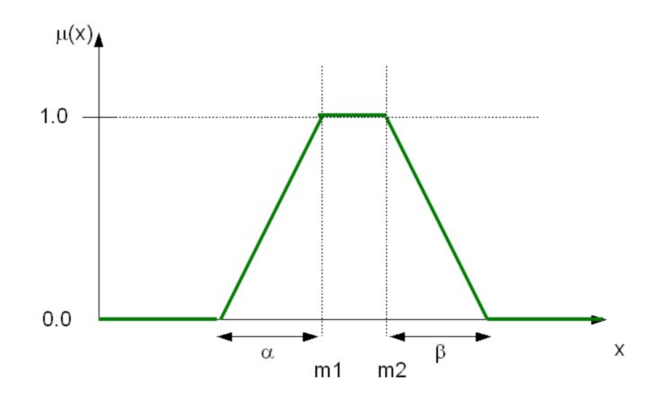


Abbildung 6.7: Fuzzy-Trapez - Funktion

6.3 Adaption von Fuzzy Logik auf unser System

Um Fuzzy-Logik auf unser System zu adaptieren, müssen geeignete **Linguistische Variablen (LV)** definiert werden.
Eine LV ist ein sprachlicher Begriff, der verschiedene Werte annehmen kann. Diese Werte heißen **linguistische Terme (LT)**.

Es gibt zunächst zwei Unterscheidungen: Linguistische Variablen (LV) für den Kundenvektor und LVs für den Finanzvektor.
Beide müssen getrennt behandelt werden, obwohl sie sich sehr ähnlich sind (wie wir gleich sehen werden).

6.3.1 Kundenvektor

Der Kundenvektor besteht aus drei Bereichen: Sicherheit, Verfügbarkeit und Rendite.

Mittels Fuzzy-Mengen können die „feinen" Bereiche des Kunden bzw. Finanzvektors dann gröberen Beschreibungen (siehe unten) zugeordnet werden, und zwar mit gewissen Toleranzen, so dass nicht immer nur ein starrer Bereich vorgegeben ist, sondern darüber hinaus auch andere Bereiche in einer ermittelten Ordnung zugeordnet werden können.

Die folgende Grobklassifizierung (Abstimmung mit der Finanzprodukte-Gruppe) wird dann auf den Vektor übertragen, so dass später Bereiche (Fuzzy Mengen) zugeordnet werden können.

- Sicherheit
  - sehr hoch
  - hoch
  - hoch / mittel

  - mittel
  - mittel / gering
  - gering
  - sehr gering

- Verfügbarkeit
  - kurzfristig [Monate]
  - mittel [Jahre]
  - langfristig [Jahre]

- Rendite (wird nicht vom Kunden angegeben (wer will keine hohe Rendite?) sondern vom System anhand Verfügbarkeit und Sicherheit berechnet
  - sehr hoch
  - hoch
  - hoch / mittel
  - mittel
  - mittel / gering
  - gering
  - sehr gering

Die linguistischen Variablen sind hierbei: Sicherheit, Verfügbarkeit und Rendite.
Die linguistischen Terme sind (am Beispiel für die LV Verfügbarkeit): kurzfristig, mittel und langfristig.

Nun haben wir die LVs definiert. Was nun noch fehlt ist die genaue Festlegung der LTs.

Sicherheit

Für die LV Sicherheit gibt es nach unserer Grobklassifizierung 7 LTs.
Folgende Prämissen sollen für die Aufstellung der LTs gegeben sein:

- Die Mengen werden als Trapeze angegeben
- Länge eines Trapezes $(m_2 - m_1)$: 9 Einheiten
- Abstand zum nächsten Trapez : 6 Einheiten
- Alpha bzw. Beta (Abfallbereich) : 12

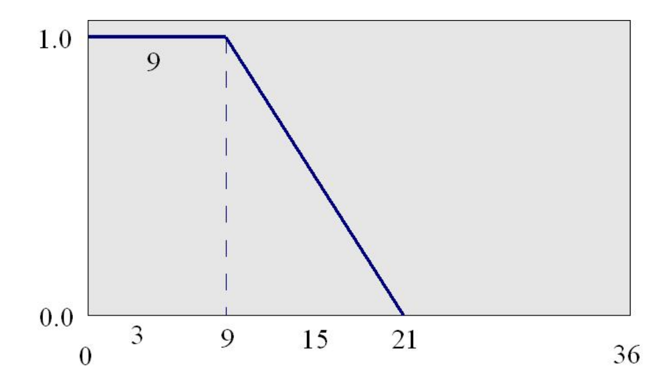


Abbildung 6.8: LT Sicherheit = ,,sehr gering"

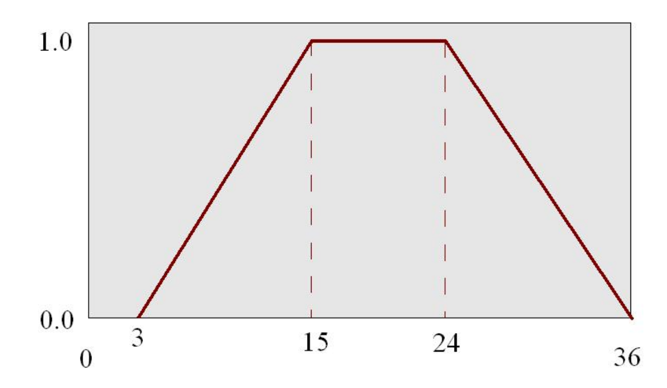


Abbildung 6.9: LT Sicherheit = ,,gering"

Hierdurch wird der gesamte Sicherheitsbereich von 0 .. 100 in die obigen 7 Sicherheitsklassen geteilt. Jeder Sicherheitsklasse wird eine einzelne Fuzzy Menge zugeordnet (z.B. A = Sicherheit ist gering, B = Sicherheit ist mittel, ..)

Die zwei Grafiken (6.8 und 6.9) sollen die Bildung der LTs etwas genauer verdeutlichen.

Insgesamt zeigt Abb. 6.10 die Darstellung der linguistischen Variablen ,,Sicherheit" mit allen untergeordneten LTs, so dass die Bildung der LTs an diesem Beispiel klar verdeutlicht wird.

Rendite

Die Fuzzy Mengen für Rendite werden genauso wie die für Sicherheit definiert. Sollten diese in weiteren Adaptionen noch in weniger als die 7 Gruppen eingeteilt werden so kann dieses mittels den Fuzzy Mengeoperatoren durch Kombination geschehen. Vorerst (in unserer Implementierung), genügen die 7 Klassen.

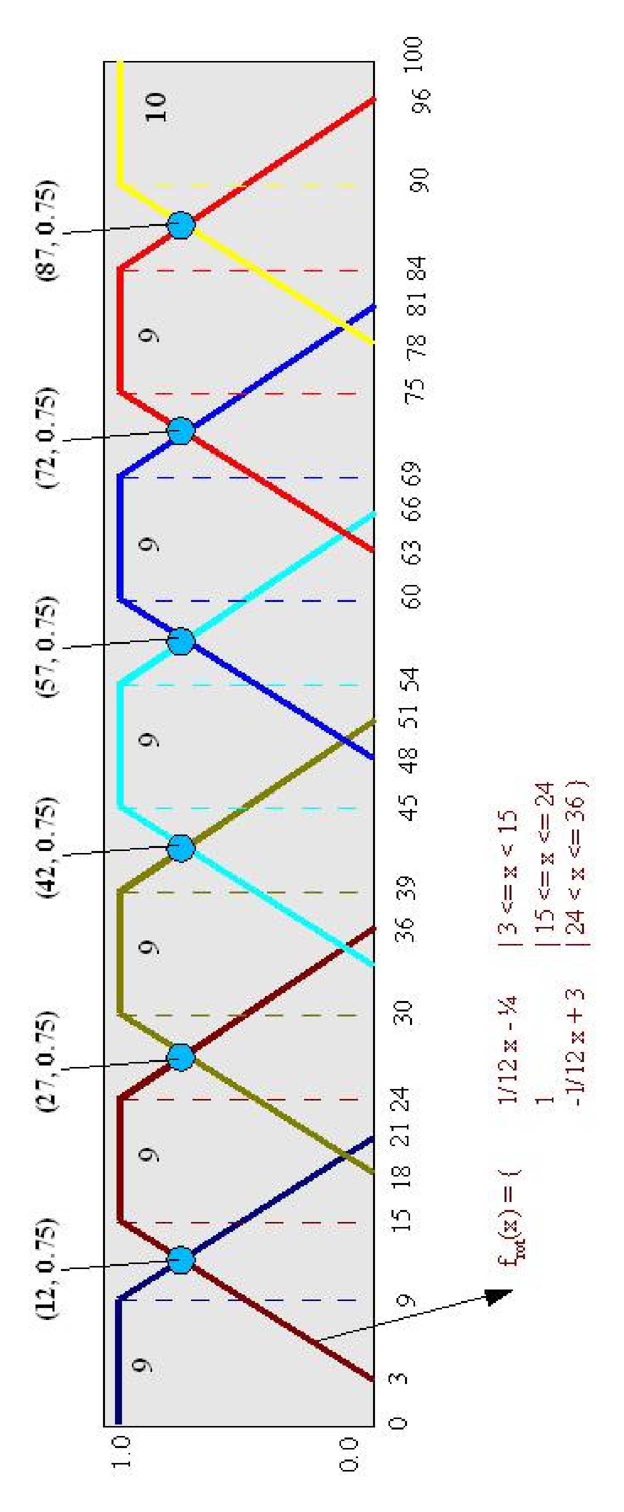


Abbildung 6.10: Darstellung der LV Sicherheit mit allen LTs

Verfügbarkeit

Die LTs für Verfügbarkeit werden scharf angegeben, da die Verfügbarkeit vom Kunden angegeben wird und somit vorerst nicht vom Programm optimiert wird. Das heißt insbesondere, dass der LT ,,kurzfristig" alle Werte zwischen 0 und 33 auf 1.0 setzt, alle anderen auf 0.0. Der LT ,,mittelfristig" setzt alle Werte von 34 bis 66 auf 1.0, alle anderen auf 0.0 und analog der LT ,,langfristig", nur dass dieser auf dem Bereich 67 bis 100 operiert.

6.3.2 Finanzvektor

Der Finanzvektor besteht - nicht wie der Kundenvektor aus drei - sondern nur noch aus zwei Bereichen: Sicherheit und Marketperformer.
Die Verfügbarkeit wird hierbei nicht mit in das System integriert. Das liegt daran, dass wir uns vornehmlich mit Aktien aus dem DAX beschäftigen wollten und somit die Verfügbarkeit für alle diese Aktien nicht mit in das System integriert werden muss, da die Verfügbarkeit für alle Aktien gleich ist.

Die folgende Grobklassifizierung wird dann auf den Vektor übertragen, so dass später Bereiche (Fuzzy Mengen) zugeordnet werden können.

Sicherheit

Die Sicherheit des Finanzproduktes wird nicht mehr durch 7, sondern nur noch durch 4 Klassen beschrieben, da wir wie gesagt nur Aktien betrachten.
Bei Aktien ist es aber so, dass keine Aktie eine hohe Sicherheit aufweisen kann, da alle Prognosen bezüglich des Verlaufs einer Aktie höchst spekulativ sind und niemand wirklich genau vorhersagen kann, wie sich der Kurs verändert. Schon kleinste Gerüchte oder Firmeneklats sind in der Lage, den Kurs drastisch zu verändern.
Von daher haben wir entschieden, dass keine Aktie eine Sicherheit von mehr als ,,mittel" in Bezug auf die oberen Klassifizierungen besitzen kann, und somit nehmen wir für die Aktie nur noch die unteren vier Klassen, welcher einer geringeren Sicherheit entsprechen.

Die 4 Klassen bestimmen sich zu:

- ,,Sicherheit sehr gering"
- ,,Sicherheit gering"
- ,,Sicherheit gering/mittel"
- ,,Sicherheit mittel"

Die Fuzzy-Mengen hierzu werden nun auch nicht mehr durch Trapeze angegeben, sondern durch Dreiecke, die wie folgt gesetzt werden:

- Sicherheit sehr gering: m $= 0$, $\alpha = 0$, $\beta = 22$.

- Sicherheit gering: m = 22, $\alpha = 22$, $\beta = 22$.
- Sicherheit gering/mittel: m = 44, $\alpha = 22$, $\beta = 22$.
- Sicherheit mittel: m = 66, $\alpha = 22$, $\beta = 22$.

Man sieht hierbei, dass durch diese Aufteilung sehr hohe Sicherheitswerte nicht mit abgedeckt werden (nach 66 sinkt die Zugehörigkeit für mittel auch schnell ab), was auch so beabsichtigt ist.

Marketperformer

Zunächst: Die zugehörige Performerklasse eines Finanzproduktes wird NICHT in Fuzzy-Logik umgesetzt, sondern kann von den Ergebnissen aus Kapitel 4 schon direkt mit quasi-Zugehörigkeitswerten angegeben werden, so dass eine Logik nicht mehr implementiert werden muss.

Die Performerklasse eines Finanzproduktes wird (nach den Ergebnissen in Kapitel 5) in drei Klassen eingeteilt, die sich zu folgenden bestimmen:

- underperformer
- marketperformer
- outperformer

In Kapitel 4 wurde bereits auch ein Schema bzw. ein Algorithmus entworfen, um den Marketperformer eines Finanzproduktes (speziell bei uns: für eine Aktie) bestimmen zu können.
Dieser Algorithmus gibt für jede der drei Klassen eine Art ,,Zugehörigkeit" zu diesen Klassen an, wobei es jedoch durch die Berechnung sein kann, dass die Zugehörigkeit für alle diese Klassen annähernd gleich wird. In so einem Fall (wenn man kein eindeutiges Maximum feststellen kann und somit die Klasse eindeutig zuordnen kann) soll der Marketperformer einer Aktie zu ,,marketperformer" bestimmt werden.

Diese Anpassung sieht dann nur noch so aus, dass man den Zugehörigkeitswert für underperformer und outperformer eines Finanzproduktes vergleicht. Sollten diese Werte weniger als 0.2 auseinanderliegen, so kann man keinen eindeutigen Schluss ziehen (da man nicht sicher sagen kann, in welche Klasse das Finanzprodukt fällt) und bestimmt somit den Marketperformer eines Finanzproduktes zu ,,marketperformer".
In dem anderen Fall wählt man die Klasse mit der höchsten Zugehörigkeit und hat somit den Marketperformer eines Finanzproduktes klar bestimmt, so dass sich nun damit weiter rechnen lässt.

6.4 Verschmelzung von FP und Kunde

6.4.1 Grundlegende Ideen

Da wir jetzt die einzelnen Fuzzy-Mengen für den Kunden und den Aktien bestimmt haben, können wir uns anschicken, den Kunden mit jeder Aktie zu verschmelzen und mit den Zugehörigkeitswerten zu den Fuzzy-Mengen vom Kunden und von der Aktie zu verknüpfen. Die Hauptidee besteht darin, den Schnitt der Fuzzy-Menge für Sicherheit von dem Kunden mit jeder Fuzzy-Menge für Sicherheit der Aktie zu berechnen und den Performer der Aktie in dem Zugehörigkeitswert der Rendite des Kunden mit einfließen zu lassen.

Konkret sieht das für den Sicherheitswert so aus: Wenn das Finanzprodukt einen positiven Zugehörigkeitswert zu einer Sicherheits Fuzzy-Menge hat, was maximal bei zweien der Fall sein kann, wird diese Menge mit der Sicherheitsmenge des Kunden, die nach rechts erweitert wird, mit Hilfe der Min-Operation geschnitten. Die Sicherheitsmenge kann deswegen getrost erweitert werden, da ein Kunde sicherlich nichts dagegegen haben wird, wenn es auch noch mögliche Aktien gibt, die eine noch höhere Sicherheitseinschätzung besitzen, da es für den Kunden auf jeden Fall laut seiner Einstellung auch ok wäre. Von den entstehenden Schnittmengen wird der Flächeninhalt ausgerechnet. Der resultierende Wert wird mit dem Zugehörigkeitswert zu der Finanz-Sicherheits-Fuzzy-Menge und der Kunden Sicherheit Fuzzy-Menge multipliziert. Dies wird für alle positiven Zugehörigkeitswerte der Finanz-Sicherheits-Fuzzy-Menge aufaddiert.

Formel: $\sum_j \mathrm{A}(K_i \;, FP_j) * \mu_{K_i} * \mu_{FP_j}$

Um nun die Rendite des Kunden bzw. Marketperformer der Aktie einfließen lassen zu können, werden die Ergebnisse der Marketperformer-Analyse "vorbearbeitet", um zu bestimmen zu welcher Klasse (out, market, under) die Aktie denn wirklich gehört. Dazu: Wenn die Differenz zwischen Out- und Under-Zugehörigkeit < 0.2 wird die Aktie als market eingeschätzt, sonst wird die Aktie zu der Klasse mit maximalem Wert zugeordnet. Das Ranking, das auf Sicherheitswerten beruht wird nun noch anhand des Marketperformers "gewichtet", um auch die Rendite mit einzubeziehen. Je nach Zugehörigkeit des FP zur Marketperformer-Einteilung wird die passende Zeile der folgenden Tabelle ausgerechnet und aufaddiert:

-	sehr gering	gering	gering/mittel	mittel	mittel/hoch	hoch	sehr hoch
under	$1 * \mu_{K_{s.g.}}$	$0{,}6 * \mu_{K_g}$	$0{,}3 * \mu_{K_{g/m}}$	$0 * \mu_{K_m}$	$0 * \mu_{K_{m/h}}$	$0 * \mu_{K_h}$	$0 * \mu_{K_{s.h.}}$
market	$1 * \mu_{K_{s.g.}}$	$1 * \mu_{K_g}$	$1 * \mu_{K_{g/m}}$	$1 * \mu_{K_m}$	$0{,}6 * \mu_{K_{m/h}}$	$0{,}3 * \mu_{K_h}$	$0 * \mu_{K_{s.h.}}$
out	$1 * \mu_{K_{s.g.}}$	$1 * \mu_{K_g}$	$1 * \mu_{K_{g/m}}$	$1 * \mu_{K_m}$	$1 * \mu_{K_{m/h}}$	$1 * \mu_{K_h}$	$1 * \mu_{K_{s.h.}}$

Diese Einteilung bedeutet, dass man den Zugehörigkeitswert entsprechend

des Performergrades gewichtet. So wird die Zugehörigkeit der jeweiligen Renditeklasse des Kunden einfach übernommen, falls die Aktie in die out-Menge fällt. Bei der market-Klasse sollte man den Wert nur dann ändern, wenn der Bereich der Rendite Fuzzymenge in mittel/hoch oder höher fällt. Wenn die Aktie eher ein Underperformer ist, ist sie nur dann interessant, wenn der Kunde eine eher geringere Rendite anstrebt.

6.4.2 Ranking

Auf den resultierenden Rankingwert gelangt man nun, wenn man den Sicherheitsendwert mit dem Wert, der bei der Verschmelzung vom Performer und Renditewert herauskommt, multipliziert. Da man für den DAX30 also 30 verschiedene Rankingeinträge erhält, ist es komfortabel eine absteigende Rankingliste auszugeben, bei der der größte Wert auch an erster Stelle steht usw...

Kapitel 7

Entscheidungsfindung (Erweiterung: Einbeziehung von News)

7.1 Einleitung

Unser Programm berücksichtigt bisher nur die Fundamentaldaten der Unternehmen und gibt auf deren Grundlage Empfehlungen zum Kauf der Aktien ab. Diese Fundamentaldaten werden allerdings nur einmal im Jahr von den Unternehmen veröffentlicht, so dass unser Programm auch folgerichtig nur einmal im Jahr den Status einer Aktie verändern würde. Dies ist natürlich nur bedingt geeignet um eine Empfehlung zum Kauf einer Aktie machen zu können, da das Programm dadurch viel zu statisch ist und nicht auf aktuelle Ereignisse und Meldungen reagieren kann. Deshalb wurde das Programm erweitert, so dass es sich die Finanzmeldungen aus dem Internet holen kann. Diese werden vom System bewertet und dann in die Bewertung der Aktie einfliessen. Bei der automatischen Bewertung der Nachrichten durch unser Programm gibt es allerdings einige Schwierigkeiten (vgl. Kapitel 6), so dass wir uns für eine andere Möglichkeit zur Bewertung der Nachrichten entschieden haben. Es wurde das sogenannte Easy-IR in das Programm eingebaut. Dabei kann der Benutzer die Nachrichten bewerten und so entscheiden, ob eine Nachricht für das Unternehmen erstens überhaupt relevant und zweitens, ob sie positiv für das Unternehmen ist. Diese Bewertungen werden dann in die Analyse des Unternehmens mit einbezogen, so dass eine wesentlich bessere, da aktuellere Einschätzung des Unternehmens möglich ist.

7.2 Lernen mithilfe des Easy-IR-Systems

7.2.1 Bewertungsabgabe eines Kunden

Der Benutzer kann sich die in der Datenbank gespeicherten Nachrichten anzeigen lassen. Entweder kann er sich die Nachrichten zu den Unternehmen, die er in seinem Portfolio hat ansehen oder er kann in einer Liste von Nachrichten sich die entsprechenden aussuchen. Sollte er sich die Nachrichten aus der Liste ansehen, so kann er bei der Bewertung angeben, für welches Unternehmen er die Einschätzung geben will, ansonsten wird ihm dies durch das Programm vorgegeben. Bei der Bewertung kann er sowohl die Relevanz, als auch die Güte der Nachricht für das Unternehmen einstellen. Dazu stellt er den für die Relevanz vorgesehenen Schieberegeler auf einen Wert zwischen 0 (für nicht relevant) und 100 (für absolut relevant). Mit dem zweiten Regeler kann er angeben, wie gut die Nachricht seiner Meinung nach für das Unternehmen ist. Hier kann er ebenfalls die Werte zwischen 0 (sehr schlecht für das Unternehmen) und 100 (sehr gut für das Unternehmen) angegeben. Sollte die Nachricht älter als 28 Tage sein, so wird diese Einschätzung nicht mehr gespeichert, da die Nachricht bereits zu alt ist und die weitere Kursentwicklung nicht mehr beeinflussen wird.

7.2.2 Bewertung des Kundenstatus

Da allerdings anzunehmen ist, dass sich nicht alle Benutzer gleich gut in der Finanzwelt auskennen und deshalb auch die Nachrichten nicht gleich gut einschätzen können, wurde eine weitere Komponente in das Programm eingefügt. Dies ist der sogenannte Kundenstatus, der angibt wie gut oder schlecht der Benutzer in der Vergangenheit die Nachrichten eingeschätzt hat und daher auch in der Zukunft besser oder schlechter geeignet ist Vorherzusagen zu treffen. Dieser vom Kunden nicht einsehbare Status ergibt sich folgendermaßen: Der Kunde gibt für die Nachricht im Bezug auf ein bestimmtes Unternehmen seine Bewertung für Relevanz und Tendenz ab. Für seine Bewertung der Nachricht steht ihm nur eine gewisse Zeitspanne zur Verfügung, so dass er nicht abwarten kann wie die Aktie auf diese Meldung reagiert und dann erst seine Einschätzung dem System mitteilt. Soll nun eine neue Empfehlung für den Kauf von Aktien gegeben werden, wird überprüft, ob die Zeitspanne zum Abgeben von Einschätzungen bereits überschritten ist. Ist dies nicht der Fall, kann das System berechnen, wie gut der Benutzer mit seiner Einschätzung lag. Dazu holt er sich den Kursverlauf der Aktie und den des DAX und schaut sich vier verschiedene Zeitpunkte "in der Näh" der Erscheinungszeit der Nachricht an. Daraus berechnet er die Kursentwicklung zwischen zwei benachbarten Zeitpunkten im Vergleich zum DAX. Ist der Aktienkurs stärker gestiegen, als der des DAX, so wird dies positiv gewertet. Entsprechend gilt, dass eine in etwa gleiche Entwicklung des Aktienkurses mit dem des DAX als neutral und eine schwächere Entwicklung als negativ gewertet wird. Je stärker nun die gegebene Einschätzung mit der berechneten Entwicklung übereinstimmt, desto größer ist der Betrag, der dem Benutzer auf seinen Kundenstatus als Bonus gutgeschrieben wird. Dabei kann ein Wert zwischen -3

und +3 erreicht werden, wobei eine -3 bedeutet, dass der Benutzer total falsch lag und eine +3 eine völlige Übereinstimmung bedeutet. Dieser Bonus wird allerdings noch mit der vom Benutzer gegebenen Relevanz (geteilt durch 100) multipliziert, bevor er zum vorherigen Kundenstatus addiert wird.

7.2.3 Berechnung der durchschnittlichen User-Tendenz

Damit das System eine Einschätzung geben kann, wie gut eine Aktie ist, werden natürlich alle von den Benutzern gemachten Einschätzungen berücksichtigt. Dazu werden diejenigen Nachrichten genommen, für die es mindestens fünf Benutzereinschätzungen für ein Unternehmen gibt. Nachrichten mit weniger als fünf Bewertungen werden aufgrund des mangelnden Interesses der Benutzer nicht mit in die Bewertung für das Unternehmen aufgenommen. Die Gesamttendenz ergibt sich nun folgendermaßen: Sollte ein Benutzer einen negativen Kundenstatus haben, so haben sich seine Vorhersagen in der Vergangenheit als nicht zutreffend erwiesen und seine Bewertung fließt daher nicht mit in die Gesamttendenz ein. Ansonsten wird das Produkt aus der vom Kunden angegebenen Tendenz, der von ihm angegebenen Relevanz und seinem Kundenstatus gebildet. Dieses wird durch die Summe aller Kundenstatus, die die Nachricht bewertet haben (und deren Kundenstatus positiv ist), geteilt und dann mit der Anzahl multipliziert. Dadurch erhält man die nach dem Kundenstatus und den gegebenen Relevanzen gewichtete Gesamttendenz dieser Nachricht für das Unternehmen. Das System bildet nun noch das nach den Relevanzen gewichtete Mittel aller Gesamttendenzen der für dieses Unternehmen dem Benutzer abgegebenen Nachrichten und erhält dadurch die Easy-IR-Einschätzung des Unternehmens aufgrund der von den Benutzern bewerteten Nachrichten.

7.2.4 Anpassung des Marketperformers einer Aktie

Damit das Programm zu einer endgültigen Einschätzung des Unternehmens gelangt, werden die Einschätzungen aus den Fundamentaldaten und aus dem Easy-IR noch verschmolzen. Dabei gewichtet das Programm die Easy-IR Einschätzung umso stärker, je mehr Bewertungen für Nachrichten es zu diesem Unternehmen von Kunden mit positiven Kundenstatus gibt. Dabei beträgt der maximale Anteil, den die Easy-IR Einschätzung an der endgültigen Einschätzung haben kann, 75 Prozent.

Die genaue Formel für den Anteil, den die Easy-IR Einschätzung an der letztendlichen Einschätzung hat, lautet:

$$\text{Anteil von Easy-IR } = \min(((\text{Anzahl der Bewertungen} - 5)/2), 75) \text{ in Prozent}$$

Diese Gewichtung wird nun wie folgt verwendet, um heraus zu finden, ob das Unternehmen ein Under-, Market- oder Outperformer ist:

$$\text{Gesamteinschätzung} = (\text{Fundamentalzahleneinschätzung} * (1 - \text{gewicht})) + (\text{Easy-IR Einschätzung} * \text{gewicht})$$

Wenn die Gesamteinschätzung kleiner als 0,66 ist, so wird das Unternehmen als Underperformer klassifiziert, ist der Wert zwischen 0,66 und 1,33, so wird das Unternehmen als Marketperformer klassifiziert, ansonsten ist es ein Outperformer.

Kapitel 8

Beschaffung benötigter Daten

8.1 Fundamentale Kennzahlen durch HTML-Wrapper

8.1.1 Beschreibung Kennzahlen

Im Kapitel über die Klassifikation von Finanzprodukten werden die fundamentalen Kennzahlen KGV, PEG, DIV, KCV, KBV, MU, CM, EBIT, EBITDA und EKR für das betrachtete Unternehmen verwendet. Dort finden sich auch die Kennzahlbeschreibungen. Um diese Daten zu erhalten, wird die Internetseite www.onvista.de abgefragt. Sie stehen auf mehreren Seiten verteilt der Öffentlichkeit zur Verfügung. Des weiteren sind die ISIN-Nummer und die zugehörige Branche von Interesse. Ein einfaches Programm schreibt diese Daten dann in die Datenbank.

8.1.2 Konzept

Der Fundamentaldatenwrapper macht sich die statische Natur der Webseite zunutze. So stehen die gesuchten Daten immer an der selben Stelle im HTML-Quellcode. Der Wrapper merkt sich lediglich die Buchstabenkombination auf die das gesuchte Datum folgt (also insgesamt 12 Zeichenfolgen für einen Datensatz). So steht beispielsweise die ISIN-Nummer immer hinter der Zeichenfolge ”ISIN:” und wird nach hinten durch das Zeichen ”<” begrenzt. Dabei muss die verwendete Zeichenfolge aber keine semantische Bedeutung besitzen. So steht der Name der Aktie hinter der nichtssagenden Zeichenfolge ”OnVista”.

Für jede Aktie existiert eine solche Webseite. Die URLs zu diesen Seiten unterscheiden sich nur durch die OnVista-interne Identifikationsnummer. Diese Nummern haben wir manuell extrahiert und in eine Liste zur automatischen Abarbeitung eingefügt. Die extrahierten Datensätze werden anschliessend mit Hilfe eines einfachen SQL-Befehls in die Datenbank eingetragen.

Die durchschnittlichen Werte für jede Kennzahl wurden per Hand berechnet und der Datenbank zugefügt.

8.1.3 Ausgabe

Es erfolgt keine für den Benutzer sichtbare Ausgabe, da die Werte nur systemintern zur Bewertung verwendet werden. Nach der Ausführung des Wrappers stehen dann die Fundamentalkennzahlen, die Branche und die ISIN aller Aktien im DAX30 zugriffsbereit in der Datenbank zur Verfügung.

8.2 Finanznachrichten durch RSS Wrapper

8.2.1 Einleitung

Der RSS-Wrapper ist eine Implementierung, um Nachrichten aus dem Internet in die FIPs Datenbank zu übertragen. Das Konzept und der Hintergrund des RSS-Wrappers wird im folgenden erläutert. Die Problemstellung Nachrichten zu bewerten bzw. zu klassifizieren beruht entschieden darauf, welche Nachrichten zugrunde gelegt werden.

Da FIPs zur Zeit nur das Finanzprodukt Aktie berücksichtigt, sind nur Nachrichten über Unternehmen interessant, die an der Börse verzeichnet und über das Internet verfügbar sind. Die für uns interessanten Nachrichten unterscheiden sich in der Struktur der Nachricht und dem Inhalt. Nachrichten aus dem Internet liegen in folgenden Formaten vor:

- Newsletter
- Feeds
- Webseiten
- Newsgroups
- Ticker
- Weblogs
- Dokumentdateien (PDF, DOC, PS)

Um die Komplexität der Nachrichtenquellen zu verringern, werden im folgenden Newsgroups, Ticker und Weblogs nicht mehr beachtet, da diese auch nicht den primären Weg darstellen über den offizielle Nachrichten von Unternehmen die Öffentlichkeit erreichen. Daraus ergeben sich folgende mögliche Formate, die bei der Extraktion von Nachrichten aus dem Internet zu beachten sind:

Format	Kurzbeschreibung
RSS	XML Format zum Austausch von Nachrichten, strukturiert
XBRL	XML Format zum Austausch von Geschäftsberichten, strukturiert
RDF	XML Format zum Austausch von Nachrichten, strukutiert
HTML	Webseiten, unstrukturiert
PDF, DOC, PS	Dokumentenformat, unstrukturiert

Hinsichtlich des Inhalts unterscheiden sich Nachrichten in:

- Unternehmensberichte
- Marktberichte
- Adhoc Meldungen
- Analysen
- Empfehlungen
- Allgemeine Nachrichten aus Magazinen, Online-Magazinen

Evaluierung der Inhalte der Nachrichten zeigt dabei folgende Probleme:

Unstrukturierte HTML Nachrichten:

Nachrichten im HTML Format liegen in einem unstrukturierten Format vor, d.h. jede Webseite zeigt die Nachrichten anders an, mal mittig in der Webseite oder von Werbebanner durchzogen oder als Block mit anderen Nachrichten. Das heißt die Trennung zwischen dem Text der Nachricht, die aus einer Webseite herausgefiltert werden soll, und dem Rest des Textes der Webseite, wie Werbung, weitere Nachrichten oder Navigationstexte, ist nie eindeutig. Die Strukturelemente von HTML geben dabei auch keine semantischen Hinweise. Für einen Nachrichten-Wrapper liegt nun die Schwierigkeit darin entweder für möglichst viele Nachrichtenwebseiten deren Struktur (wo steht jetzt die Nachricht) zu kennen und entsprechend zu extrahieren oder anhand von Merkmalen, die einen Nachrichtentext auszeichnen automatisch ohne Wissen über die zugrunde liegende Webseite den Inhalt zu extrahieren.

Unternehmensberichte:

Unternehmensberichte für Unternehmenskenndaten sind fast ausschließlich im PDF Format in einem erzählenden Text vorhanden. Das bedeutet, die für uns wichtigen Kenngrößen sind daraus eigentlich nicht automatisch berechenbar, da sich in solchen Berichten bunte Charts und Tortendiagramme mit “um den heißen Brei“ geschriebenen Text abwechseln. Der eigentlich erhoffte Standard XBRL ist im Internet z.Z. jedenfalls noch nicht so verbreitet für die Öffentlichkeit zugänglich. Wahrscheinlich wird dieser Standard überwiegend intern oder auf nicht öffentlichen Kommunikationswegen genutzt (Unternehmen -

Finanzamt). Die Deutsche Börse stellt zwar einige Berichte zur Verfügung, diese umfassen allerdings zu wenige und für uns uninteressante Unternehmen.

Daraus resultiert, dass hier der Schwerpunkt auf strukturierte Nachrichten gesetzt wird. Als strukturierte Nachrichten bieten sich die RSS-Feeds an. RSS-Feeds liegen in einem standardisierten Format vor und eignen sich für die maschinelle Weiterverarbeitung, da sie auf dem XML-Format basieren.

```
<?xml version="1.0" encoding="iso-8859-1"?>
<rss version="">
  <channel>
    <title>Titel des News-Feeds</title>
    <link>Link zur Webpräsenz</link>
    <description>Kurzbeschreibung</description>
    <language>de</language>
    <copyright>Copyright-Informationen</copyright>
    <pubDate>Datum der Erstellung</pubDate>
    <image>
      <url>URL des Bildes für eine Darstellung des Bildes</url>
      <title>Titel des Bildes </title>
      <link>Allgemeiner Link zum Bild</link>
    </image>
    <item>
      <title>Titel des ersten Beitrages</title>
      <description>Der Text der News</description>
      <link>direkter Link zu weiterführenden Angaben der News</link>
    </item>
     ...
  </channel>
</rss>
```

Dabei gibt die Kurzbeschreibung schon eine kurze Zusammenfassung um was es in dieser Nachricht eigentlich geht. Das Problem bei RSS-Feeds besteht aber weiter darin, dass die vollständige Nachricht erst über den Link im Feed selber abrufbar ist und dieser Link halt wieder auf eine Webseite verweist, welche die Nachricht entsprechend des Webseitenlayouts der referenzierten Seite darstellt (siehe Probleme mit Unstrukturierte HTML Nachrichten). Nun gibt es RSS-Feeds von bestimmten Anbietern (hier am Beispiel Finanztreff.de) wo die weiterführenden Links auch auf die Seite von Finanztreff.de verweisen. Die Nachrichten also auf Seiten verweisen, von denen der RSS-Feed stammt. Bei anderen RSS-Feeds ist es allerdings so, dass sie unterschiedliche RSS-Feeds in einen zusammenführen und somit die weiterführenden Links auf vollkommen unterschiedliche Webseiten verweisen. Es gibt nun wie oben beschrieben zwei Möglichkeiten für die Implementierung des Wrappers: Wrapper mit Wissen über die Struktur jeder referenzierten Webseite oder Wrapper die anhand von Merk-

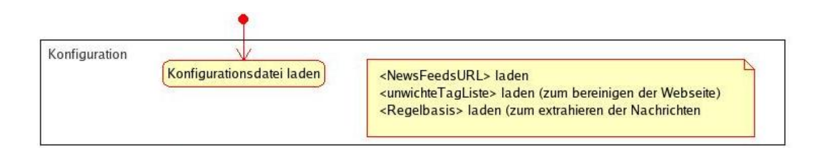


Abbildung 8.1: RSS: Konfigurationsdatei lesen

malen von Nachrichten diese aus Webseiten extrahiert (ohne Wissen über die Struktur der referenzierten Webseite)

Für FIPs wird die zweite Möglichkeit implementiert. Die Grundidee dabei ist, die Kurzbeschreibung einer Nachricht im RSS-Feed zu nutzen und die Wörter dieser Kurzbeschreibung in der entsprechenden verlinkten Webseite zu suchen. Eine Webseite kann in der Regeln als DOM-Baum dargestellt werden. Da ein längerer zusammenhängender Text oft innerhalb eines *TD,P,DIV* HTML-Elements vorkommt, kann der Wrapper den Zweig im Baum identifizieren, wo die meisten ähnlichen Wörter der Kurzbeschreibung vorkommen und diesen dann als Nachricht extrahieren. Dieser Wrapper ist weitgehend webseitenunabhängig. Der Text wird dann bereinigt um vorhandene HTML-Tags (Image-Links, Anchor etc, bold, center) und als Nachricht interpretiert.

8.2.2 Konzept

Der Funktionsablauf des RSS-Wrappers gliedert sich in 6 Schritte.

Im ersten Schritt wird die Konfigurationsdatei gelesen. Diese beinhaltet folgende für den RSS-Wrapper relevanten Einstellungen (siehe Abbildung 8.1):

Eigenschaft	Kurzbeschreibung
MinTitleEqualsPercent	Wieviel Prozent vom RSS Titel reichen aus, um den Titel auf der Seite des Deep Links zu identifizieren ?
MinTextEquals	Wieviele Wörter reichen aus, um einen Text zur RSS Beschreibung zuordnen zu können ?
MinTextWords	Wievele Wörter muss eine Textphase haben, um als Textphase interpretiert werden zu können ?
MinTextPhrases	Wieviele Sätze muss eine Textphase haben, um als Textphase interpretiert werden zu können ?
MaxEmptyTextTagsBetweenText	Wieviele nicht zu berücksichtigen Textphasen dürfen maximal zwischen identifizierten Textphasen liegen ?
Feeds	Liste von RSS-Feed URLs

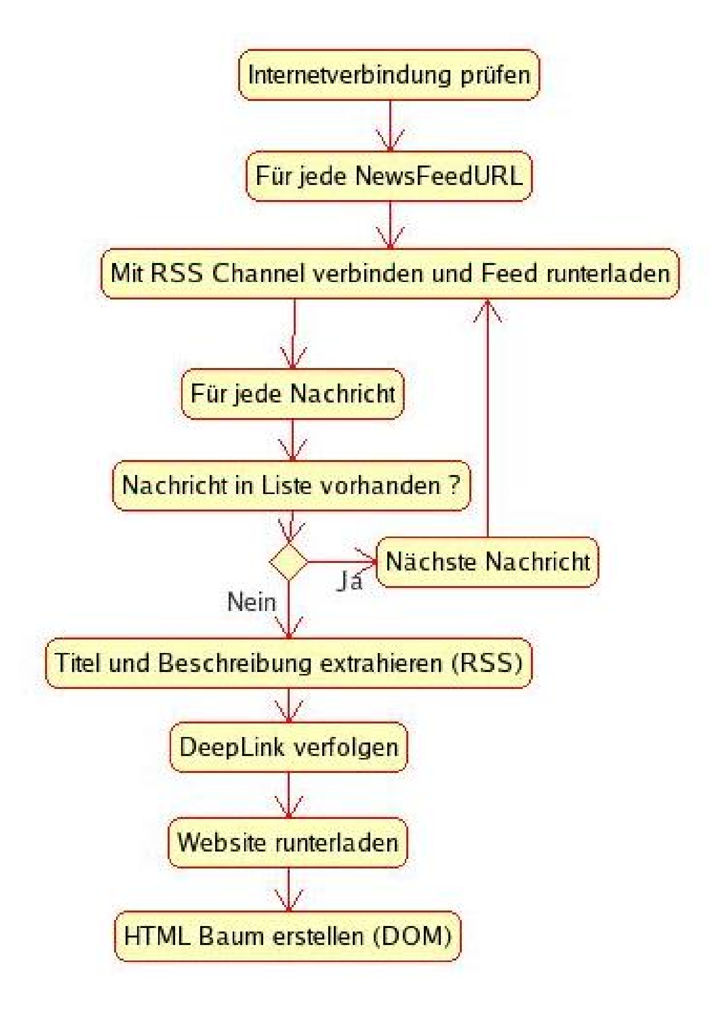


Abbildung 8.2: RSS: HTML Code extrahieren und DOM-Baum erstellen

Im zweiten Schritt werden für jeden RSS-Feed die Nachrichteninformationen wie Titel, Beschreibung und URL der eigentlichen Nachricht (DeepLink) extrahiert (siehe Abbildung 8.2). Für jede Nachricht eines Feeds wird dann die URL der Webseite, auf die die Nachricht referenziert, heruntergeladen und in einen DOM-Baum überführt. Die Überführung in einen DOM-Baum setzt dabei Wohlgeformtheit der HTML-Seite voraus. Dies ist u.U. nicht gegeben, deshalb wird mittels der OpenSource Bibliothek JTidy (http://jtidy.sourceforge.net/) der HTML Code der runtergeladenen Webseite zuvor in wohlgeformtes HTML überführt.

Im dritten Schritt werden alle unwichtigen Elemente (Tags) aus dem DOM-

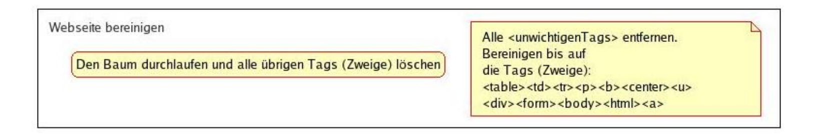


Abbildung 8.3: RSS: DOM bereinigen

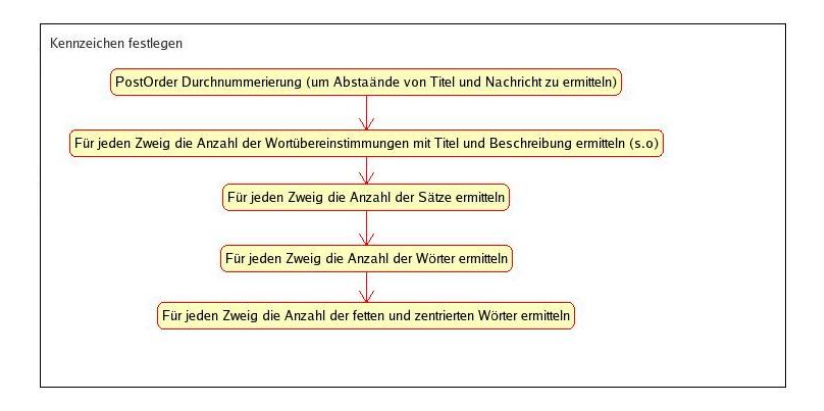


Abbildung 8.4: RSS: Element Merkmale berechnen

Baum entfernt, damit nur noch Elemente übrigbleiben, die potentiell auch Text enthalten können (siehe Abbildung 8.3).

Im vierten Schritt werden für jeden Zweig (jedes Element des DOM-Baumes) die folgenden Kennzeichen ermittelt. Die Kennzeichen werden im fünften Schritt dazu verwendet, um anhand der in der Konfigurationsdatei festgelegten Regeln zu bestimmten, ob in einem HTML Element die relevante (oder Teile der relevanten) Nachricht steht (siehe Abbildung 8.4).

Im fünften Schritt wird aufgrund der folgenden Regelbasis ein Zweig (Element) als relevant für eine Nachricht angesehen oder auch nicht (siehe Abbildung 8.5).

1. Wenn mindestens der Text eines Elementes mit *MinTitleEqualsPercent* mit dem Titel der Nachricht aus dem RSS-Feed übereinstimmt. Diese Regel sorgt dafür, dass mit großer Wahrscheinlichkeit der Anfang der Nachricht gefunden wird.

2. Wenn der Text eines Elementes mindestens *MinTextEquals* Wortübereinstimmungen mit der Beschreibung der Nachricht aus dem RSS-Feed hat.

3. Wenn der Text eines Elementes mindestens aus *MinTextWords* Wörtern

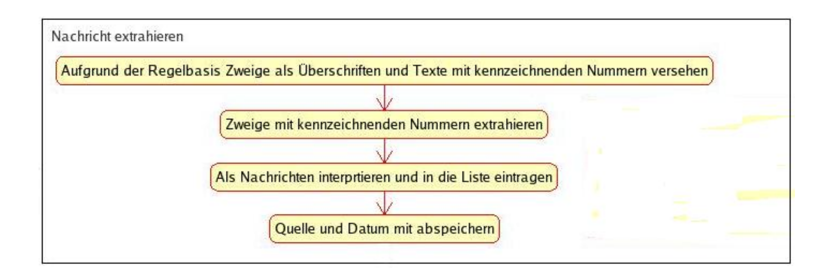


Abbildung 8.5: RSS: Regelbasis anwenden

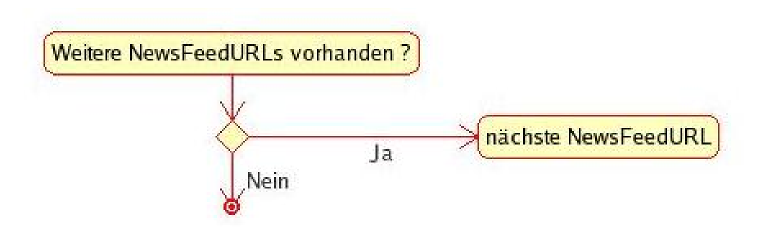


Abbildung 8.6: RSS: Über Nachrichten Iterieren

besteht

4. Wenn der Text eines Elementes mindestens aus *MinTextPhrases* Sätzen besteht

5. Wenn zwischen zwei Elementen (wobei das erste Element schon relevant für eine Nachricht ist) maximal *MaxEmptyTextTagsBetweenText* nicht relevante Elemente stehen. Damit wird sichergetsellt, nachdem der Anfang (Titel) einer Nachricht gefunden worden ist, dass die folgenden Elemente auch Nachfolger (Textpassagen die zur Nachricht gehören) der Überschrift sind. Keine direkten Nachfolger, da Texte häufig durch Werbeelemente unterbrochen sind, aber auch nicht zu weit entfernte Nachfolger.

In Schritt 6 wird über die Nachrichten und RSS-Feeds iteriert (siehe Abbildung 8.6).

8.2.3 Ausgabe

Die extrahierten Daten werden in der FIPs Datenbank in die Tabelle Nachricht geschrieben. Dabei wird als Datum der Nachricht das aktuelle Datum genommen, wann die Nachricht in die Datenbank geschrieben wird. Zunächst ist der Ansatz gewesen, das Datum der Nachricht zu nehmen wie es im RSS-Feed definiert ist, doch viele RSS-Feeds pflegen diese Angabe nicht. Die Implementierung

des RSS-Wrappers ist ohne grafische Benutzeroberfläche und Benutzerinteraktion. Somit wurde der Aufruf des RSS-Wrappers mittels einem CRON-Job auf dem PG-Server angestoßen, so dass jeden Tag einmal alle Nachrichten aus den in der Konfigurationsdatei angegebenen RSS-Feeds geladen und in die Datenbank eingetragen werden.

8.3 SPAM-Filter

8.3.1 Konzept

Einleitung: Da in unseren Nachrichten, die vom NewsWrapper gezogen werden, viele Nachrichten enthalten sind, die gar nichts über Aktien behandeln, kam die Idee auf, einen Spamfilter zu implementieren, der sich diese irrelevanten News schnappt. So interessiert sicherlich wenig, ob gerade in der Bundesliga Bayern gegen Dortmund spielt oder flix gegen flax. Die Schwierigkeit liegt darin, dass die Nachrichten meistens nicht ganz so eindeutig zu trennen sind, wie es z.B. bei E-Mails der Fall ist durch Filterung z.B. des Wortes SSex". Ein häufiger Ansatz besteht darin, eine Support Vektor Maschine (SVM) mit Daten zu füttern und die SVM dann selber versucht eine Hyperebene zwischen relevanten und irrelevaten Nachrichten zu finden.

SVM: Dieser Ansatz sah in erster Linie viel versprechend aus, da auch in Literaturrecherchen SVM immer benannt werden, wenn es in irgendeiner Art und Weise um Spamfilterung geht. So klassifizierten wir gut 500 Stück der Nachrichten als relevant (-1) bzw. irrelevant (1), damit wir schon einmal einige Testdaten zur Verfügung hatten. Die gaben wir dann der Gruppe SSVMäls Eingabe für ihr Programm, das auch klassifizieren können sollte.

Ergebnis SVM: Das Ergebnis war allerdings nicht zufrieden stellend, da nur knapp 70% der Nachrichten richtig als Spam identifiziert wurden. Es scheint, dass SVM besser geeignet gewesen wären, wenn die Nachrichten inhaltlich stärker getrennt wären.
Dies war natürlich keinesfalls ein Grund den Kopf in den Sand zu stecken, so dass wir nach weiteren Möglichkeiten gesucht haben.

BayesClassifier: Nach weiterer Recherchearbeit stießen wir auf ein OpenSource Projekt namens BayesClassifier, das auch im Netz angepriesen wurde und unter www.sourceforge.net frei verfügbar ist. Es benutzt, wie der Name schon sagt, Bayessche Wahrscheinlichkeiten, um eine neue Nachricht zu klassifizieren. Das Programm funktioniert grob folgendermaßen: Als Input erhält es eine Menge von Daten, die schon vorklassifiziert sind. Mit Hilfe von Bayes werden dann Wahrscheinlichkeiten für Wörter berechnet, die

ausschlaggebend dafür sind, dass diese Nachricht gejunkt bzw. relevant ist. Das gute und einfache ist, dass man dem Programm kein Wörterbuch zur Verfügung stellen muss, sondern es für sich ein ëigenes"Wörterbuch mit niedrigen Wahrscheinlichkeiten für Spamwörter und hohen Wahrscheinlichkeiten für Nicht-Spam-Wörter erstellt. Die Formel für die Wahrscheinlichkeitsberechnung einer neuen Nachricht sieht wie folgt aus:

Formel vom BayesClassifier: $Prob(spam \mid words) = \frac{Prob(words|spam)*Prob(spam)}{Prob(words)}$

Diese Wahrscheinlichkeiten wendet das Programm auf neue Nachrichten an und versucht dann anhand der Wortwahrscheinlichkeiten richtig zu kategorisieren. Zu dem benötigt man natürlich noch Stoppwortlisten die dafür sorgen, dass Artikel usw. nicht in die Wortwahrscheinlichkeiten aufgenommen werden.

Also dachten wir, geben wir ihm viele Trainingsdaten (ca. 1000) zum Berechnen der Wahrscheinlichkeiten und 500 Testdaten, anhand derer wir unsere Ergebnis verifizieren konnten. Die Spamerkennung endete in dem Bereich von 40%. Also bei weitem schlechter als SVMs.

Doch da kam uns eine Verbesserung in den Sinn, die den BayesClassifier besser machen sollte. Wir gaben ihm schon bestimmte Worte vor, die darauf hinwiesen, dass die Nachricht Spam war bzw. Relevanz aufzeigte.

Dafür verwendeten wir sogenannte Blacklists, eine für "guteünd eine für ßchlechte"Nachrichten. Wenn das Programm jetzt ein Wort findet, dass in einer Hamlist vorkommt, ist die Nachricht relevant, wohingegen ein Vorfinden eines Wortes aus der Spamlist darauf hindeutet, dass es sich um eine Spam-Nachricht handelt. Um auch mögliche Nachrichten klassifizieren zu können, wo Worte aus beiden Listen vorkommen, verwendet der BayesClassifier ein Wahrscheinlichkeitmaß, dass angibt, wie er die Nachricht einzuordnen hat. Bei einem Wert von 0 bis 0,5 erkennt er die News als relevant an, sonst bei dem Intervall größer 0,5 bis 1 als Spam. Man muss die Worte natürlich mit Bedacht wählen und mögliche Doppeldeutigkeiten zwischen diesen Listen ausschließen, also wenn das Wort "WM"vorkommt, dann wird es Spam sein, wobei man so ein Wort wie Ëntwicklung"nicht generell einer bestimmten Liste zuordnen kann. Beide Listen können vom User verwaltet und modifiziert werden, so dass es auch möglich ist, bestimmte Bereiche komplett auszublenden.

8.3.2 Ausgabe

Dieses Vorgehen führte dazu, dass der BayesClassifier zu 80% Junks richtig erkennt, was schon eine deutliche Verbesserung zu dem eigentlichen BayesClas-

sifier und SVMs ist.

Ein weiteres Ziel vom SpamFilter war es, doppelte und leere Nachrichten aus der Datenbank zu entfernen bzw. gar nicht erst in der Datenbank abzulegen. Dafür verwendeten wir einen Algorithmus (md5) der einen Message Digit für jede Nachricht erzeugt, so dass komplett syntaktisch ähnliche Nachrichten mit Überprüfung auf Identität mit dem Wert gefunden und nicht gespeichert werden.

8.4 Wörterbuch

8.4.1 Konzept

Um eine bessere Kategorisierung der Nachrichten zu erreichen, als über bloßes Clustering nach der Häufigkeit aller Wörter, ist es nötig, relevante Wörter zu bestimmen, die als Bestimmungsgrundlage für CI- oder KI-Verfahren dienen. Da so ein Finanzwörterbuch noch nicht vorhanden war, haben wir selbst ein Wörterbuch erstellt. Das Wörterbuch wurde mittels einer graphischen Oberfläche erstellt. Dort wurden Finanznachrichten aus der Datenbank dargestellt. Die PG-Teilnehmer markierten Wörter, die sie als relevant für die Einschätzung der Nachricht empfanden; diese wurden auf ihre Stammform reduziert und im Wörterbuch gespeichert. Außerdem konnte man ausgewählte Wörter zu Synonymklassen zusammen fassen oder ihnen Antonyme zuweisen.

8.4.2 Ausgabe

Das fertige Wörterbuch wurde als Relation mit Synonymen in der Datenbank (Tabelle Woerterbuch) gespeichert. Hier ein Auszug aus den insgesamt 361 gestemmten Begriffen in der Datenbanktabelle: gewinn (mit synonym profit) vereinbar börsenaufsicht notier beschäftigt mitarbeit gesenkt Ökonom minus verkauf erlos gesellschaft kapitalerhoh kapital steig (mit Synonym zugelegt) investition zertifikat mitglied spekulativ teilt jahresabschluss festgestellt rucklag aufsichtsrat reduziert information vorwurf schad zinszahl

8.5 Testdaten

Da in der FIPs Datenbank alle möglichen Nachrichten zu allen möglichen Unternehmen stehen, die selbt nach dem sogenannten Spam-Filter so nicht für Testzwecke zu gebrauchen sind, wird eine Methode gebraucht, um geeignete Testnachrichten zu erhalten. Diese wird in folgendem Abschnitt beschrieben.

Für eine bessere Kontrolle der Daten wäre es gut, diese von Hand einzugeben und zu bearbeiten. Da aber für die meisten der getesteten Verfahren eine große Menge solche Nachrichten verfügbar sein muß und diese ein realistisches

Bild der Nachrichtenlage widerspiegeln sollen, wird auf eine automatische Vorfilterung und Vorverarbeitung der Nachrichten der FIP Datenbank gesetzt.

Weiter werden für überwachte Verfahren bereits klassifizierte Daten benötigt, daher ist auch eine automatische Klassifikationsmethode gesucht.

8.5.1 Bearbeitungspipeline der Nachrichten

- **Filterung der FIP Datenbank:** Es werden alle Nachrichtentexte der FIP-DB gewählt, in denen ein oder mehrere Unternehmens-Suchwörtern vorkommen. Diese Suchwörter werden vorher festgelegt und jedem Wort das zugehörige Aktienzeichen zugeordnet.
  Bsp: *basf* → *BAS.DE*, *b.a.s.f.* → *BAS.DE*.
- **Speichern der Nachrichtentexte:** Gefundene Nachrichtentexte werden nun mit dem zugehörigen Aktienzeichen in der Datenbank *tesdaten* gespeichert. Kommen in einer Nachricht verschiedene Unternehmen vor, so wird die Nachricht jeweils nur dem ersten Unternehmen zugeordnet. Dies soll verhindern, dass Nachrichtentexte öfter als einmal in der DB vorkommen, was zum Beispiel bei der Trennung in eine unabhängige Trainings- und Testliste von Beispielnachrichten zu Problemen führen würde.
- **Preprocessing der einzelnen Nachrichten**
  1. **Lemmatisierung:** Der Nachrichtentext wird mit Hilfe des TreeTaggers (siehe 5.4.2) lemmatisiert.
  2. **Stoppwörter entfernen:** Dananch wird der lemmatisierte Text mit Hilfe einer Stoppwortliste um Stoppwörter bereinigt. Dieser Schritt kann nicht dem TreeTagger vorausgehen, da dieser für das POS-Tagging auch Stoppwörter benutzt.
  3. **Synonymgruppen finden:** Zu jedem Wort wird nun mit Hilfe des OpenThesaurus die Synonymgruppe bestimmt. Dabei wird für jede gefundene Synonymgruppe ihr *fipscounter* um eins erhöht. Damit wird gezählt, wie oft diese Klasse insgesammt in *testnachrichten* vorkommt, um später daraus ein einfaches Wörterbuch zu erstellen.

  $\rightsquigarrow$ **Lemma der Form:**
  (Typ Wort w_1) Synonymgruppe $s_{1,1}$ / ... / Synonymgruppe s_{1,l_1}, ... ,
  (Typ Wort w_n) Synonymgruppe $s_{n,1}$ / ... / Synonymgruppe s_{n,l_n}
  Beispiel: (VVFIN)3459/7259,(NN)12480,($.),(ADJD)6031/10527/13794/14804/4556

  Schließlich wird das Lemma zur Nachricht in *testdaten* gespeichert.

8.5.2 Berechnung der Bewertung einer Nachricht

- **Holen der Aktienkurse:** Zu den betrachteten Aktien (entspicht den Aktienzeichen der Unternehmens-Suchwörter) werden anhand des Aktienzeichens die Börsenkurse der letzten 200 Tage von Yahoo!Finance geholt.

Weiter werden die Kurse des DAX im gleichen Zeitraum geladen. Diese Daten dienen als Grundlage zur Berechnung der Nachrichtenbewertungen.

- **Berechnung der Nachrichten-Bewertungen:** Anhand der Kursverhältnisse Aktie zu DAX werden (wie unter Abschnitt 5.5.7 beschrieben) die Bewertungen der einzelnen Nachrichten berechnet und in der *testdaten* DB gespeichert.

8.5.3 Erstellen eines einfachen Wörterbuchs

Nachdem alle Testnachrichten vorverarbeitet sind kann anhand des *fipscounters* abgelesen werden, welche Synonymgruppen besonders oft in Nachrichtentexten vorkommen und welche nicht oder nur selten auftreten.

Die Idee ist nun ein einfaches Wörterbuch aus Synonymklassen zu erstellen, die oft, aber nicht zu oft vorkommen. Zu häufig verwendetet Wörter haben keine große Aussagekraft, zu seltene Wörter sind eben wegen der geringen Überdeckung ungeeignet. Es wird also ein einfaches Wörterbuch aller Synonymgruppen erstellt, für die gilt $min \leq$ fipscouter $\leq max$. Die Wahl von min und max ist dabei abhängig von der Anzahl betrachteter Nachrichten und der gewünschten Größe (Strenge) des Wörterbuchs.

Kapitel 9

Speicherung der Daten im System

Die Datenbank bildet unterste Schicht des FIPs-Systems. Zur Persistenz der Daten wird das Postgre-SQL-Datenbanksystem verwendet. Die Daten werden in einer relationalen Datenbankstruktur gespeichert. Der Zugriff von Java aus auf die Datenbank erfolgt mittels SQL-Anweisungen über die JDBC-Schnittstelle, die seit Version 1.4 Bestandteil der Java-2-Plattform ist.

9.1 DB-Schema

Die beiden folgenden Abbildungen zeigen das Konzept, auf dem die Datenbank arbeitet. Dabei bildet das erste Diagramm nur die Relationen und deren Beziehungen zueinander ab. Das zweite Bild führt die Attribute der einzelnen Relationen auf.

Im Zentrum des Datenbankkonzepts befindet sich die Tabelle *Aktie*. Eine Aktie wird identifiziert durch ihre ISIN. Zusätzlich werden der Name des Unternehmens zu dem die Aktie gehört und der Link zu der Homepage des Unternehmens abgespeichert. Des Weiteren werden alle Fundamentalkennzahlen wie *peg*, *kbv*, *kgv*, die zu der Aktie gehören abgespeichert. Diese Daten sind die Werte aus dem letzten Jahr.

Jede Aktie gehört zu einer Branche. Auch alle Bezeichnungen für die Fundamentalkennzahlen werden in einer Tabelle *Kennzahl* abgespeichert. Die Tabelle *Branchendurchschnitt* stellt eine Beziehung zwischen Kennzahl und Branche dar. Aus ihr kann der Durchschnitt einer Fundamentalkennzahl bezüglich einer Branche berechnet werden.

Die Tabelle Nachricht besitzt neben *titel*, *datum*, *text*, *autor* und *quelle* die beiden Attribute *gatenewsid* und *gatecorpusid*. Diese beiden Attribute beinhalten die Identifikationswerte, die einer Nachricht von Gate vergeben werden. *Spamwert* und *shavalue* beschreiben, inwieweit eine Nachricht sich auf Aktien bezieht oder eine Spam-Nachricht darstellt.

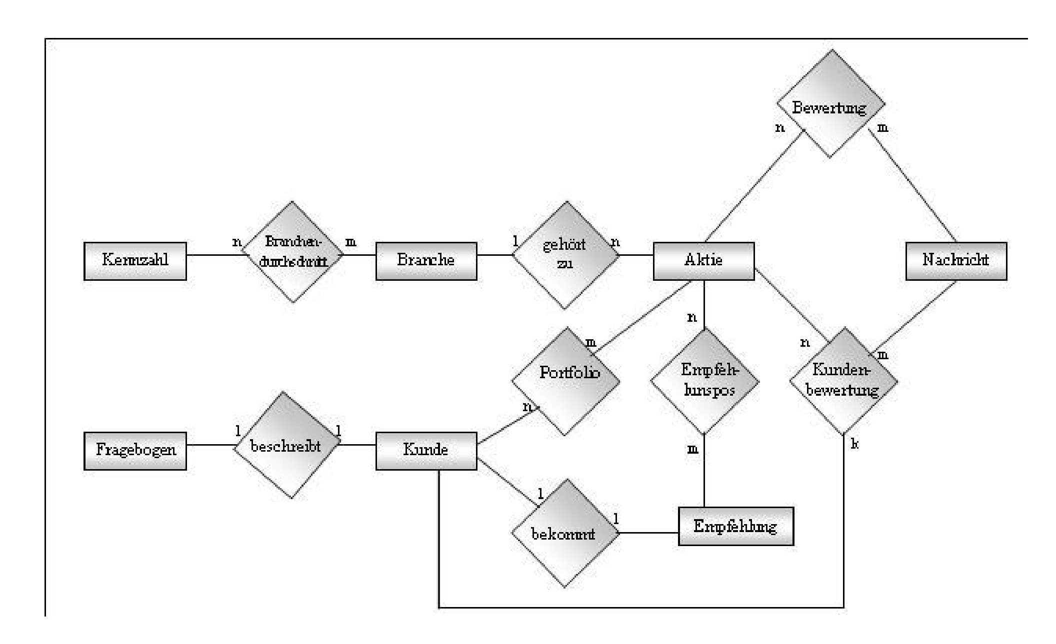


Abbildung 9.1: Datenbankschema

Die Tabellen *Aktie* und *Nachricht* werden über zwei Beziehungstabellen miteinander verbunden: *Bewertung* und *Kundenbewertung*. Die Tabelle *Bewertung* gibt an, welche Nachrichten das System von sich aus welchen Aktien zugeordnet hat. Wenn möglich, soll in dieser Tabelle auch die Entscheidung des Systems darüber gespeichert werden, inwieweit eine Nachricht relevant für die Bewertung einer Aktie ist und welche Tendenz diese Bewertung hat. *Relevanz* und *tendenz* stellen Integer-Zahlen, wobei eine hohe Tendenz eine positive Bewertung bedeutet. Bei der Beziehungstabelle *Kundenbewertung* handelt es sich um die von Kunden gegebenen Relevanz- und Tendenzwerte. Diese Daten entstehen, indem die Kunden auf der GUI des Systems an einer Skala von 1-100 festlegen, welche Relevanz und Tendenz eine Nachricht bezüglich einer Aktie aufweist. Dadurch, dass das Datum einer Kundenbewertung festgehalten wird, können bei der Erstellung von Aktien-Rankings stets die aktuellsten Kundenbewertungen herangezogen werden. Das Attribut *bonusgegeben* wird gebraucht bei der späteren Beurteilung darüber, ob eine Kundenbewertung auch tatsächlich der Entwicklung der Aktie entsprochen hat.

Eine weitere zentrale Tabele im Datenbankkonzept ist die Tabelle *Kunde*. Neben der eindeutigen *kundennr*, mit der sich der Kunde im System einloggt, beinhaltet diese Tabelle die Attribute *sicherheit*, *verfuegbarkeit* und *rendite*. Diese Werte geben die Erwartungen und das Anlageverhalten des Kunden wieder und werden aus den Informationen ermittelt, die der Kunde als Antworten in den Fragebogen eingibt. Das Attribut *kundenstatus* sagt aus, inwieweit der Kunde verlässliche Bewertungen abgegeben hat. Die Bewertungen eines Kunden mit einem hohen kundenstatus werden bei der Erstellung eines Aktien-Rankings

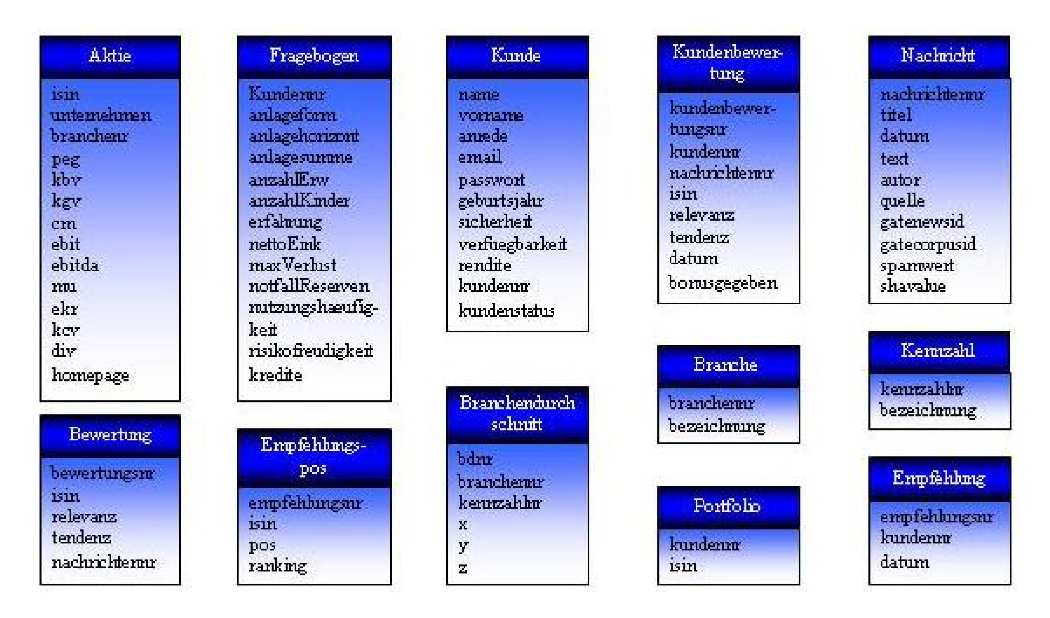


Abbildung 9.2: Attribute der Datenbankrelationen

stärker in Betracht gezogen.

Die Tabelle *Fragebogen* enthält alle wichtigen Informationen, die aus den Antworten mit denen ein Kunde bei der Registrierung im System den Fragebogen ausfüllt, erschließbar sind. Diese Daten können vom Kunden während der Nutzung des Systems verändert werden; dementsprechend verändert sich natürlich auch das Kundenprofil, das sich aus den Attributen *sicherheit*, *verfuegbarkeit* und *rendite* der Tabelle *Kunde* zusammensetzt.

Die Tabelle *Portfolio* stellt eine Beziehung zwischen *Kunde* und *Aktie* dar und beinhaltet für einen Kunden Referenzen auf alle Aktien, die in seinem aktuellen Portfolio sind.

Die Tabellen *Empfehlung* und *Empfehlungsposition* beinhalten alle wichtigen Informationen zu einem Aktein-Ranking, das vom System dem Kunden dargeboten wird. *Empfehlung* zählt alle Empfehlungen auf, die für einen Kunden gemacht wurden. Dabei wird jede Empfehlung durch eine Empfehlungsnummer identifiziert und das Datum, an dem die Empfehlung erstellt wurde mitgespeichert. Aus der Tabelle *Empfehlungspos* kann dann für jede Empfehlungsnummer entnommen werden, welche Aktien in der Empfehlung enthalten sind und welche Position und welchen Rankingwert die einzelnen Aktien in der entsprechenden Empfehlung besitzen.

9.2 Datenbank Klassen

Die Datenbankklassen sind in zwei Pakete gegliedert: *common* und *dbcontroller*. Das Paket *dbcontroller* ist für den Zugriff auf die Datenbank zuständig und

das Paket *common* beinhaltet Klassen deren Objekete für den Datenaustausch zwischen der Datenbank und den anderen Systemkomponenten dienen.

9.2.1 Das Paket *common*

Die Klassen im Paket *common* dienen im Wesentlichen dazu, die Informationen die von der Datenbank abgefragt werden in kompakter Form in Objekte gekapselt zu den höher liegenden Komponenten des Systems zu schicken. Ausser bei den Klassen *Ranking* und *Rankingeintrag* entsprechen die Klassen weitesgehend den Tabellen im Datenbankschema. Ein Objekt der Klasse *Rankingeintrag* stellt eine Aktie innerhalb einer Empfehlung für einen Kunden dar. Die Klasse *Ranking* verwaltet eine verkettete Liste, die alle Aktein einer Empfehlung beinhaltet. Da es für den Datenaustausch zwischen den Komponenten nicht erforderlich ist und es die Handhabung der Objekte komplizierter machen würde, wurden Assoziationen zwischen den Klassen weitesgehend vermieden.

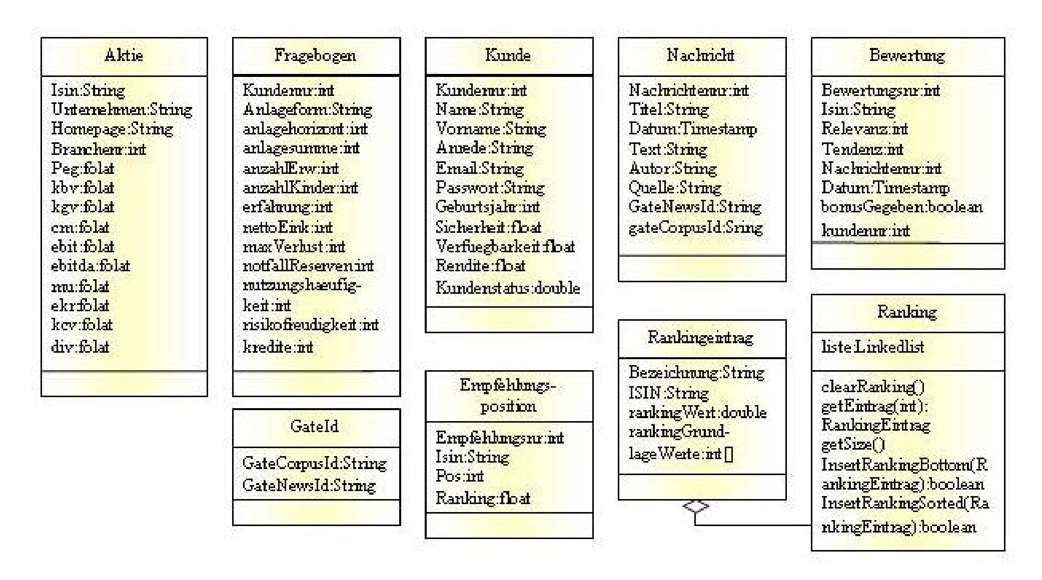


Abbildung 9.3: Klassen im Paket common

9.2.2 Das Paket *dbcontroller*

Im Paket *dbcontroller* befinden sich die beiden wichtigen Klassen *DBcontrol* und *Statements*. In der Klasse *DBcontrol* wird der Aufbau der Verbindung zur Datenbank und der Zugriff auf die Daten realisiert. Im Konstruktor dieser Klasse wird eine XML-Datei geladen, die aus allen notwendigen Informationen besteht, die zum Aufbau einer Datenbankverbindung notwendig sind. Die Methode *initDBConnection()* baut eine Datenbankverbindung auf und speichert sie in dem Attribut *private java.sql.Connection db*. Die Methoden dieser Klasse, die dem

Abruf oder der Manipulation von Daten dienen arbeiten dann auf dieser Datenbankverbindung.

Die Klasse *Statements* beinhaltet alle SQL-Anweisungen, die von der Klasse *DBcontrol* genutzt werden. Diese Anweisungen sind in der Klasse *Statements* in Form von statischen Attributen vom Typ *String* aufgelistet. Eine solche Trennung verbessert den Überblick und die Wartung des Systems.

9.3 Zugriff

Die JDBC-Schnittstelle von Java ermöglicht uns den Zugriff auf unsere PostgreSQL-Datenbank. Mit dieser Schnittstelle ist es möglich, durch SQL-Anweisungen, die in den Java-Code gestreut sind, Daten aus der Datenbank zu holen und zu manipulieren. Es werden im Wesentlichen die unten beispielhaft aufgeführten Schritte vollzogen:

- Laden des Treibers

  Class.forName(org.postgresql.Driver);

- Aufbau der Verbindung

  Connection con;

  String url = jdbc:postgresql://pg473.cs.uni-dortmund.de/FIP

  con = DriverManager.getConnection(url, pg473, pg473);

- Erzeugen einer Anfrage

  String query = SELECT name FROM Kunde WHERE kundennr=123

  Statement stmt = con.createStatement();

  ResultSet rs = stmt.executeQuery(query);

- Freigeben der Ressourcen

  stmt.close();

  rs.close();

  con.close();

Kapitel 10

Tests

Im Folgenden sind die wichtigsten Funktionstests aufgeführt

Zu testende Funktion	Neuen Benutzer anlegen /registrieren/einloggen/Fragebogendetails einsehen/abmelden
Eingabe	Herr Anton Tester, Hauptstrasse 1, 12345 Burgstadt, anton@tester.de, 1960
Sollausgabe	Der neue User Anton Tester ist angelegt; man kann sich einloggen; man kann die Fragebogendetails einsehen; man kann sich ausloggen
Istausgabe	Der neue User Anton Tester ist angelegt (Kundennummer 36), einloggen mit Usernummer und Passwort funktioniert; die korrekten Fragebogendetails werden angezeigt; mit dem Abmelden Button gelangt man wieder zu Startseite

Zu testende Funktion	Mit falschem Passwort anmelden
Eingabe	Kundennummer: 36, Passwort: Test
Sollausgabe	Man kann sich nicht einloggen
Istausgabe	Man kann sich nicht einloggen, es wird eine Fehlermeldung angezeigt, dass das Passwort inkorrekt ist

Zu testende Funktion	Fragebogendetails ändern
Eingabe	Jahreseinkommen auf 120000 ändern; Änderung übernehmen
Sollausgabe	Jahreseinkommen ist auf 120000 geändert
Istausgabe	Jahreseinkommen ist auf 120000 geändert

Zu testende Funktion	Mein Portfolio ändern; Aktie hinzufügen
Eingabe	BASF Aktie und Telekom Aktie
Sollausgabe	BASF Aktie und Telekom Aktie werden in 'Mein Portfolio' aufgenommen
Istausgabe	BASF Aktie und Telekom Aktie sind in 'Mein Portfolio' aufgenommen

Zu testende Funktion	Mein Portfolio ändern; Aktie entfernen
Eingabe	BASF Aktie
Sollausgabe	BASF Aktie wird aus 'Mein Portfolio' gelöscht
Istausgabe	BASF Aktie wird aus 'Mein Portfolio' gelöscht

Zu testende Funktion	Meine Klassifizierung einsehen
Eingabe	-
Sollausgabe	Kundennummer, Rendite, Verfügbarkeit, Sicherheit werden angezeigt
Istausgabe	Kundennummer, Rendite, Verfügbarkeit, Sicherheit werden angezeigt

Zu testende Funktion	Persönliche Daten ändern
Eingabe	Vorname auf Antona ändern; Änderungen übernehmen
Sollausgabe	Vorname ist auf Antona geändert
Istausgabe	Vorname ist auf Antona geändert

Zu testende Funktion	Mein Portfolio
Eingabe	-
Sollausgabe	Mein Portfolio wird angeigt mit den aktuellen Aktien, dem Link zum Unternehmen und den Link zu den News zum Unternehmen
Istausgabe	Mein Portfolio wird angeigt mit den aktuellen Aktien, dem Link zum Unternehmen und den Link zu den News zum Unternehmen

Zu testende Funktion	News zum Unternehmen anzeigen
Eingabe	-
Sollausgabe	Zum Unternehmen werden die Nachrichten im Zeitraum 'News der letzen 60 Tage' anzeigen
Istausgabe	Zum Unternehmen werden die Nachrichten im Zeitraum 'News der letzen 60 Tage' anzeigen

Zu testende Funktion	News bewerten
Eingabe	
Sollausgabe	
Istausgabe	

Zu testende Funktion	Empfehlungen berechnen
Eingabe	-
Sollausgabe	Es wird eine Empfehlungsseite über die empfohlenen Aktien angezeigt auf Grundlage der Fragebogendetails und des aktuellen Portfolios
Istausgabe	Es wird eine Empfehlungsseite über die empfohlenen Aktien angezeigt auf Grundlage der Fragebogendetails und des aktuellen Portfolios

Kapitel 11

Arbeiten mit dem System

11.1 Das Finanz Informations Portal

FIPs ist ein webbasiertes Informationsportal, mit welchem sich der User Finanznachrichten anschauen und diese bewerten kann. Er bekommt aufgrund der Bewertung aller User Empfehlungen zu allen Aktien des Dax. Zudem kann der User sich speziell nur Nachrichten zu den Aktien aus seinem Portfolio anzeigen lassen bzw. Nachrichten zu den für ihn interessante Unternehmen.

11.2 Möglichkeiten für den Benutzer und Typische Abläufe

Zum Programm gelangt man über die Webseite:
http://pg473.cs.uni-dortmund.de:8080/fips/

Wenn man noch keinen Useraccount besitzt, kann man sich neu anmelden. Dies geschieht über den Link auf der Startseite "Dann melden Sie sich hier an". Über diese Seite kann sich ein bereits registrierter User auch mit seiner Kundennummer und seinem Passwort einloggen.

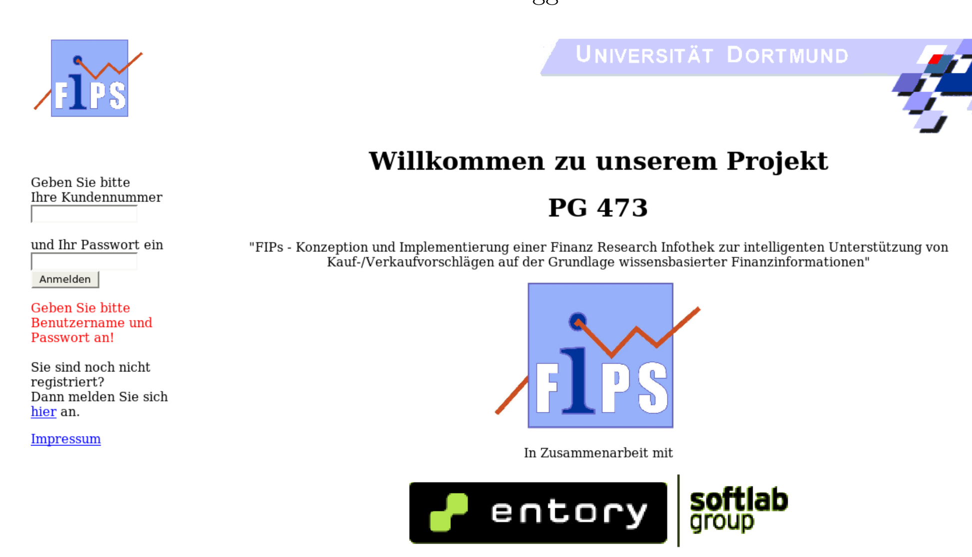


Abbildung 11.1: Login Screen

Bei der Neuanmeldung gelangt man dann zu einer Seite, in der man nach seinen für die Registrierung notwendigen Daten, insbesondere aber nach einem Passwort gefragt wird.

Abbildung 11.2: Persönliche Daten

Mit dem Button 'Speichern' gelangt man zum Fragebogen, wo der User seine finanzielle Lage, seine Anlagesumme und -horizont, sowie Erfahrungen angeben kann.



Abbildung 11.3: Fragebogen erster Teil

fIPS

UNIVERSITÄT DORTMUND

Geben Sie bitte
Ihre Kundennummer

und Ihr Passwort ein

Anmelden

Geben Sie bitte
Benutzername und
Passwort an!

Sie sind noch nicht
registriert?
Dann melden Sie sich
hier an.

Impressum

	nein teilweise
Wie lange wollen Sie Ihr Geld festlegen??	kurzfristig 1/2 bis 1 Jahr 1-2 Jahre 2-5 Jahre 5-10 Jahre >10 Jahre
Wieviel Erfahrung mit Wertpapieren haben Sie?	keine < 2 Jahre 2-5 Jahre 5-10 Jahre > 10 Jahre
Wie häufig wollen Sie unser System durchschnittlich benutzen?	täglich mehrmals wöchentlich mehrmals monatlich seltener
Wie schätzen Sie Ihre Risikofreudigkeit für die geplante Anlage ein? (Zwischen 0 und 100, wobei 100 sehr risikobereit darstellt)	35
In welche Anlageformen investieren Sie derzeit schon? (Mehrfachnennung möglich)	keine Aktien Anleihen Immobilien Devisen Sparbuch Fonds Derivate Rohstoffe andere

Speichern

Abbildung 11.4: Fragebogen zweiter Teil

Nach Betätigung des Buttons 'Speichern' gelangt man zur Bestätigungsseite, auf der man seine Kundennummer und einen Link zur Startseite erhält.



Abbildung 11.5: Anmeldebestätigung

Nun kann sich der User einloggen und sein Portfolio über den Link 'Mein Portfolio ändern' füllen. Dort kann man zunächst über den Button 'Blättern' eine Aktie aus dem DAX suchen und dann seinem Portfolio die 'Aktie hinzufügen'. Ebenfalls kann man dort mit dem Button 'Aktien entfernen' eine Aktie aus seinem Portfolio herauslöschen.





zur Verfügung stehende Aktien

Hallo Herr Mustermensch

Mein Porfolio
News bewerten

Einstellungen
Persönliche Daten ändern
Fragebogendetails ändern
Mein Portfolio ändern
Meine Klassifizierung einsehen
Abmelden

Externe Links
Finanztreff
OnVista
Finanznachrichten

Name der Aktie	ISIN	Übernehmen
ALLIANZ AG	DE0008404005	übernehmen
ALTANA AG	DE0007600801	übernehmen
BASF AG	DE0005151005	übernehmen
BAYER AG	DE0005752000	übernehmen
BAYERISCHE HYPO- UND VEREINSBANK AG	DE0008022005	übernehmen
BMW GROUP AG	DE0005190003	übernehmen
COMMERZBANK AG	DE0008032004	übernehmen
CONTINENTAL AG	DE0005439004	übernehmen
DEUTSCHE BANK AG	DE0005140008	übernehmen
DEUTSCHE LUFTHANSA AG	DE0008232125	übernehmen
DEUTSCHE POST AG	DE0005552004	übernehmen
DEUTSCHE TELEKOM AG	DE0005557508	übernehmen
E.ON AG	DE0007614406	übernehmen
FRESENIUS MEDICAL CARE AG	DE0005785802	übernehmen
HENKEL KGAA	DE0006048432	übernehmen
INFINEON TECHNOLOGIES AG	DE0006231004	übernehmen
LINDE AG	DE0006483001	übernehmen
MAN AG	DE0005937007	übernehmen
MUENCHENER RUECKVERSICHERUNGS-GESELLSCHAFT AG	DE0008430026	übernehmen
RWE AG	DE0007037129	übernehmen
SCHERING AG	DE0007172009	übernehmen
SIEMENS AG	DE0007236101	übernehmen
THYSSENKRUPP AG	DE0007500001	übernehmen

Abbildung 11.6: Aktienkatalog





Hallo Herr Mustermensch

Mein Porfolio
News bewerten

Einstellungen
Persönliche Daten ändern
Fragebogendetails ändern
Mein Portfolio ändern
Meine Klassifizierung einsehen
Abmelden

Externe Links
Finanztreff
OnVista
Finanznachrichten

Portfolio ändern

Aktien aus dem Portfolio löschen

	Name der Aktie	ISIN
○	ADIDAS-SALOMON AG	DE0005003404
○	DAIMLERCHRYSLER AG	DE0007100000
○	DEUTSCHE BOERSE AG	DE0005810055
○	SAP AG	DE0007164600

Aktie entfernen

Aktien zum Portfolio hinzufügen

Name der Aktie	ISIN
METRO	DE0007257503 Blättern

Aktie hinzufügen

Abbildung 11.7: Portfolio ändern

Über den Link 'Mein Portfolio' gelangt man zu einer Übersicht seines aktuellen Portfolios. Dort hat man die Möglichkeit, sich eine 'Empfehlung berechnen' zu lassen. 'Meine letzte Empfehlung' gibt noch einmal eine Übersicht über die Empfehlung, die bei der letzten Berechnung erstellt worden ist.



UNIVERSITÄT DORTMUND



Hallo Herr Mustermensch

Mein Porfolio
News bewerten

Einstellungen
Persönliche Daten ändern
Fragebogendetails ändern
Mein Portfolio ändern
Meine Klassifizierung einsehen
Abmelden

Externe Links
Finanztreff
OnVista
Finanznachrichten

Mein Portfolio

Name der Aktie	ISIN	durchs. User-Bewertung	Link zum Unternehmen	News zum Unternehmen
ADIDAS-SALOMON AG	DE0005003404	nicht genug Bewertungen	www.adidas-group.com	News
DAIMLERCHRYSLER AG	DE0007100000	nicht genug Bewertungen	www.daimlerchrysler.com/dccom	News
DEUTSCHE BOERSE AG	DE0005810055	nicht genug Bewertungen	www.exchange.de	News
METRO AG	DE0007257503	nicht genug Bewertungen	www.metrogroup.de	News
SAP AG	DE0007164600	nicht genug Bewertungen	www.sap.com	News

Empfehlung berechnen

Letzte Empfehlung anzeigen

Abbildung 11.8: Mein Portfolio

In der 'Mein Portfolio' Übersicht gibt es sowohl Links zu den Unternehmen in dem Portfolio, als auch einen Link, der den User zu den News bringt, die speziell zu diesem Unternehmen passen. Dort findet der User auch eine durchschnittliche Bewertung der News, wie sie von anderen Usern bereits abegegeben worden sind.



Titel	durchs. User-Tendenz	Link
Xetra: Sehr fest - Commerzbank legen deutlich zu	nicht genug Bewertungen	Zur Nachricht
Europas Börsen gut behauptet erwartet	nicht genug Bewertungen	Zur Nachricht
Infineon will Speicher-Mehrheit nach Börsengang vorerst behalten	nicht genug Bewertungen	Zur Nachricht
Infineon reduziert Verlust im Schlussquartal	nicht genug Bewertungen	Zur Nachricht
Xetra: Fester - DAX auf neuem Jahreshoch	nicht genug Bewertungen	Zur Nachricht
Xetra: Fester - DAX auf neuem Jahreshoch	nicht genug Bewertungen	Zur Nachricht
Xetra: DAX etwas fester - Zinsen verderben die Laune	nicht genug Bewertungen	Zur Nachricht

Abbildung 11.9: Newsübersicht

Per Link 'Zur Nachricht' gelangt der User zu einer Seite, auf der er bewerten kann, inwiefern der Text tatsächlich relevant für das Unternehmen ist und ob der Text sich eher positiv, negativ oder neutral auf das Unternehmen auswirkt.

FIPS

UNIVERSITÄT DORTMUND

Hallo Herr Mustermensch

Mein Porfolio
News bewerten

Einstellungen
Persönliche Daten ändern
Fragebogendetails ändern
Mein Portfolio ändern
Meine Klassifizierung einsehen
Abmelden

Externe Links
Finanztreff
OnVista
Finanznachrichten

Nachricht im Detail ansehen

Titel	Infineon schlittert weiter in die Krise
Quelle	http://www.zdnet.de/news/business/0,39023142,39140362,00.htm
Datum	2005-12-23 00:00:00.0
Autor	null
Text	Der Chiphersteller Infineon hat seine Verluste im ersten Quartal des laufenden Geschäftsjahres, das am 31. Dezember zu Ende gegangen ist, fast verdoppelt. Bei einem Umsatz von 1,7 Milliarden Euro beläuft sich das Ergebnis vor Steuern und Zinsen (EBIT) auf ein Minus von 122 Millionen Euro. Grund für das schlechte Abschneiden sei der starke Preisverfall in der Speicherchipsparte, teilt das Unternehmen mit. Der Nettoverlust liegt mit 183 Millionen Euro deutlich über dem des Vorjahres von 100 Millionen Euro.

Nachricht bewerten

ISIN der Aktie, für welche die News bewertet wird	DE0006231004 Blättern
Beurteilung der Relevanz dieser Nachricht für die Aktie [0..100] (100 = höchst relevant für Unternehmen, 0 = gar nicht relevant für Unternehmen)	100
Beurteilung der Tendenz dieser Nachricht für die Aktie [0..100] (100 = höchst positiv für Unternehmen, 0 = höchst negativ für Unternehmen)	1

Speichern

Abbildung 11.10: Nachricht im Detail

Der Link 'News bewerten' führt zu einer Übersicht der aktuellsten Nachrichten, die man ebenfalls bewerten kann und zusätzlich kann man die Nachricht einem bestimmten Unternehmen zuordnen. Auch hier sieht man eine bisherige durchschnittliche Userbewertung. Die Detailansicht entspricht der Abbildung 'Nachricht im Detail', mit dem kleinen Unterschied, dass die Möglichkeit zur Zuordnung zum Unternehmen möglich ist.



Abbildung 11.11: Nachrichten bewerten - Überblick

'Meine Klassifizierung einsehen' gibt eine Übersicht über die aktuelle Klassifizierung des Users in die Klassen 'Sicherheit', 'Rendite' und 'Verfügbarkeit'.



UNIVERSITÄT DORTMUND

Einsehen unserer Einschätzung von Ihrem Anlageprofil

Hallo Herr Mustermensch

Mein Porfolio
News bewerten

Einstellungen
Persönliche Daten ändern
Fragebogendetails ändern
Mein Portfolio ändern
Meine Klassifizierung einsehen
Abmelden

Externe Links
Finanztreff
OnVista
Finanznachrichten

Ihre Kundennummer (login) lautet : 35

Ihre Klassifikationsdaten lauten wie folgt :

Sicherheit : 19 %

Verfügbarkeit : 42 %

Rendite : 70 %

Abbildung 11.12: Klassifizierung des Users

Über 'Fragebogendetails ändern' und 'Persönliche Daten ändern' kann man bei Bedarf diese Daten aktualisieren. Die 'Externen Links' führen zu bekannten Finanzseiten im Internet. 'Abmelden' beendet die aktuelle Sitzung und man gelangt wieder auf die Startseite. Dort führt der 'Impressum'-Link zu eben diesem mit einem Gruppenbild der Teilnehmer der PG 473.

Kapitel 12

Endworte

12.1 Fazit

Welche der Ziele haben wir nun erreicht, die wir uns zu Beginn der PG gesteckt haben? Und bei welchen haben wir Schwierigkeiten gehabt?

Wir haben nach unseren Überlegungen eine Architektur für FIPs entworfen und diese weitestgehend umgesetzt. Der User kann sich eine Auswahl an Nachrichten zu Unternehmen aus seinem Portfolio anzeigen lassen oder aber auch zu solchen Unternehmen, die für ihn interessant sind. Bei den Entscheidungskomponente hat es viele Schwierigkeiten gegeben, die insbesondere auf fehlendes Finanzwissen und schwierige, und daher fehlende Einbeziehung der Semantik zurückzuführen sind. Daher arbeitet die Entscheidungskomponente nicht voll automatisch, sondern wird durch den User durch eigene Abgabe von Relevanz der Texte für ein Unternehmen und die Tendenz dieser Nachricht unterstützt. FIPs ist über ein Webinterface für den Anwender nutzbar und zudem von den Administratoren einstell- und wartbar. Der Anwender kann sich einen Useraccount anlegen und Angaben über seine finanzielle Lage und seine Vorstellung über die geplante Finanzanlage machen, die FIPs in seine Empfehlungsberechnung einbezieht. Die Erfassung dieser Daten erfolgt über einen Fragebogen. Die Aktie als einziges Finanzprodukt wird dem User sowohl aufgrund der Fundamentalkennziffern der Unternehmen, als auch durch die Bewertung der News durch die Nutzer empfohlen.

Die Muss- und ein Großteil der Wunschkriterien sind erfüllt worden.

12.2 Ausblick

Da die Bewertungskomponente nicht vollständig automatisch arbeitet, ist hier auf jeden Fall noch Raum für Erweiterungen. Besonders Verfahren, welche die Semantik von Texten analysieren, können vielleicht zu verbesserten Ergebnissen führen. Die graphische Gestaltung des Web-Interfaces kann ebenso verbessert werden, wie auch die Bedienbarkeit der einzelnen Funktionen.

Kapitel 13

Anhang

13.1 Sitzungsprotokolle

13.1.1 Protokolle 1. Semester

Sitzungsprotokoll vom 11.04.2005

Abwesend: Nils (entschuldigt)

Verspätet: Christoph (30 min), Markus (20 min)

Sitzungsleitung: Stefan Berlik

Protokollführung: Bertram

Tagesordnung

1. Begrüßung
2. Formalia
3. Seminarfahrt
   Fragen / Anregungen
   Schriftliche Ausarbeitungen
   Abrechnung s
4. PG Kasse
5. Technischer Beauftragte
6. Sonstiges
7. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt.
Protokoll führt jeder PG Teilnehmer alternierend in alphabetischer Reihenfolge.
Sitzungsleiter wird jeweils der Protokollführer des vorigen Treffens.

Seminarfahrt

- **Schriftliche Ausarbeitungen**
  Abstimmung über das gemeinsame Format der Seminarausarbeitungen:

  1x Word, 8x LaTex, 2x Enthaltung

  damit beschlossen: Ausarbeitung in LaTeX

  Abgabe: Montag, den 02. Mai 2005

- Abrechnung verschoben auf Donnerstag, den 14. April 2005

PG Kasse

- Abstimmung ob eine PG Kasse eingeführt wird

  5x dafür, keiner dagegen, 6x Enthaltung

  damit beschlossen: Einrichtung einer PG Kasse

  Einmalige Einzahlung von 10 Euro

- Festlegung der Kassenwärter und Verpflegungsmanager:

  René, Christian, Mehmet

Technischer Beauftragte

- Vorstellung des Netzwerkbeauftragten Wolfgang Hunscher
- Einführung in die Rechner des Rechner-Pools, Wolfgang Hunscher
- Vergabe der Rechneraccounts
- Festlegung des Technischen Beauftragten: Bertram

Sonstiges

- Einigung auf Erstellung eines Verzeichnis mit Namen, E-Mail Adressen, Telefon-Nummern, ICQ-Nummern aller PG Teilnehmer

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 3. Januar 2006

Abwesend: Niels (entschuldigt)

Verspätet: Ahmet (13 Min.)

Sitzungsleitung: Bertram Bödeker

Protokollführung: Jana Ehlers

Tagesordnung

1. Begrüßung
2. Formalia
3. diverse HW/SW-Szenarien a. Entwicklungsumgebung b. CVS c. Tools (Funktion, Einordnung, Methoden) d. Web-Auftritt
4. Konzeptskizze FIPs
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen. Die Rechnung von Nordhellen ist noch nicht bei Stefan angekommen, daher wird Abrechnung verschoben. ICQ-Nummern bitte noch an Stefan mailen!

diverse HW/SW-Szenarien

Diskussion darüber, ob nur englische oder nur deutsche Texte in das System eingespeist werden sollen; noch keine Einigung.

•

5 Entwicklungsumgebung Einigung auf Eclipse, jeder arbeitet sich selbst ein.

5 CVS Diskussion, ob CVS oder eine andere Sourcecodeverwaltung. Christoph arbeitet sich ein und stellt die Optionen vor.

5 Tools Datenbank: Einigung auf MySQL, Martin macht Einführung. Textmining: Bertram sucht noch nach Alternativen zu GATE, wählt dann aus und stellt vor. Tracking-/Datenverwaltungs-/Projektmanagementsystem: Madan arbeitet sich ein und stellt die Optionen vor. Design: Einigung auf togetherJ. Konzeptentwurf: Einigung auf UML.

5 Web-Auftritt Webseite der PG, auf der sie sich vorstellt und Zugriff auf Datenbank bereitstellt, wird von René, Christian und Markus verwaltet. Später ist außerdem Benutzeroberfläche für Kunden des fertigen FIPs-Systems zu erstellen. Entscheidung darüber, was und wer, wird verschoben.

Konzeptskizze FIPs

grobe Idee des Systems: Was in der Blackbox in der Mitte zu passieren hat,

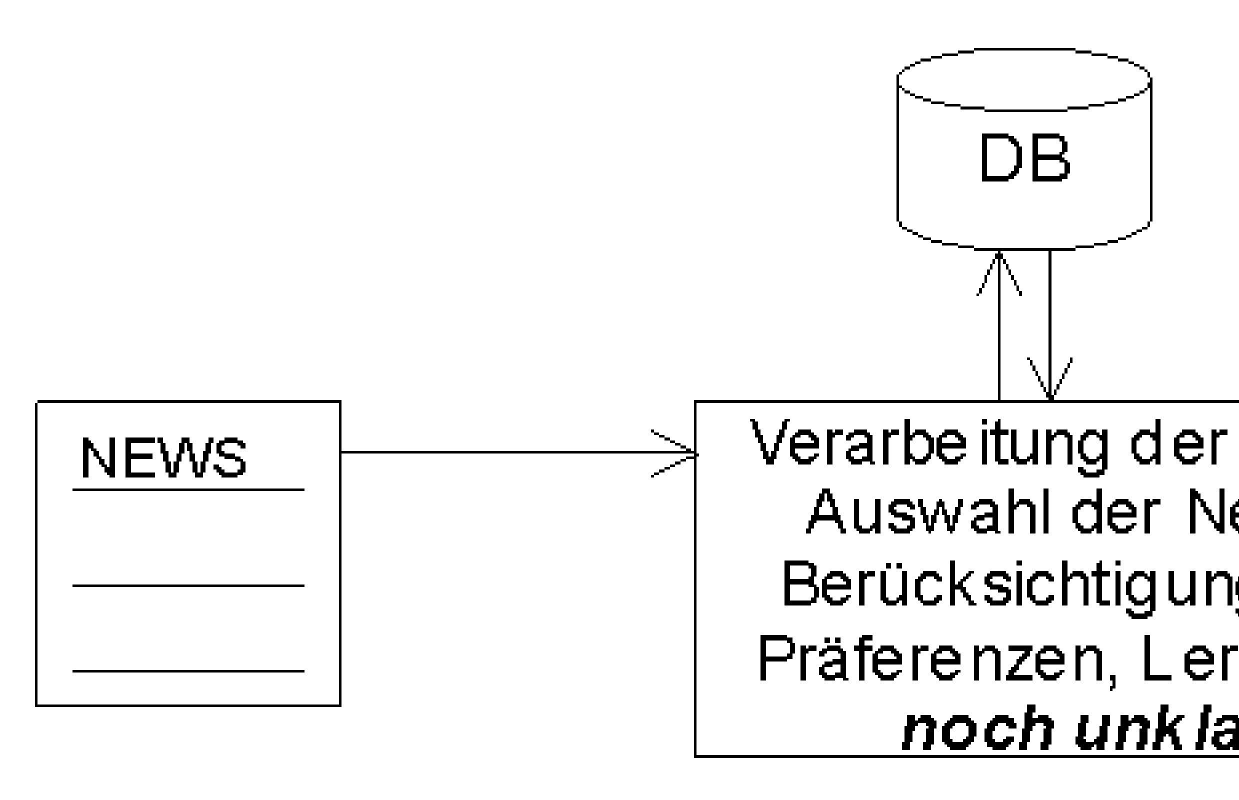


Abbildung 13.1: FIPs Grobkonzept

ist noch unklar. Bis zur nächsten Sitzung am 18.4.05 sollte sich jeder dazu Gedanken machen und eine mindmap erstellen.

Sonstiges

Schlüssel für den Pool werden verteilt. Martin gibt eine Einführung in TeX/LaTeX. Ahmed wird die TeX-Vorlage für die schriftlichen Ausarbeitungen rummailen.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 18. April 2005

Abwesend: niemand

Verspätet: Christian (10 Min.), Markus (10 Min.)

Sitzungsleitung: Jana

Protokollführung: René

Tagesordnung

1. Begrüßung
2. Formalia
3. Vorstellung, Diskussion und Fusion der verschiedenen mindmaps
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Noch immer nichts Neues über die Rechnung

Erinnerung: ICQ-Nummern an Stefan mailen!

Die Vorträge, die in der letzten Sitzung beschlossen wurden, werden am nächsten Montag, 25.04., gehalten.

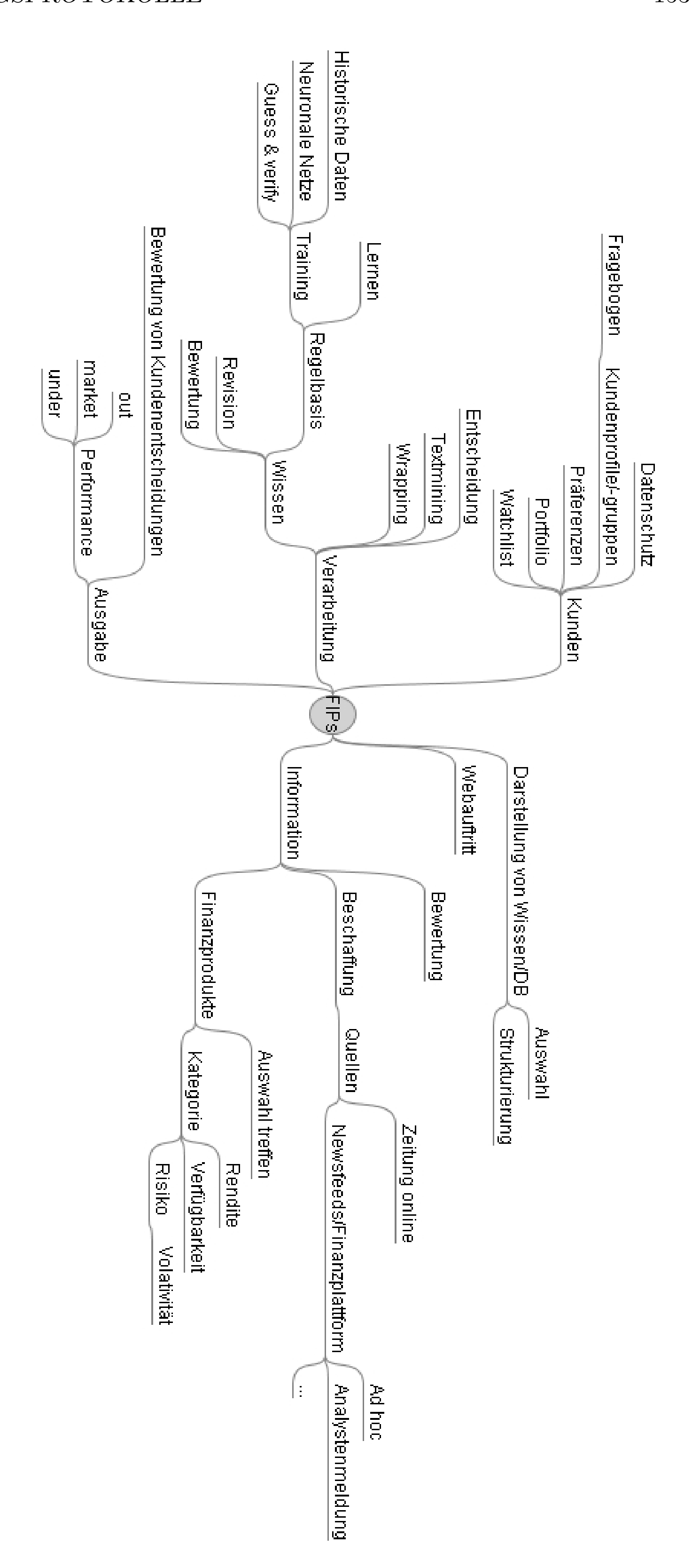
Vorstellung, Diskussion und Fusion der verschiedenen mindmaps

siehe Grafik 'FIPs Mindmap'

Sonstiges

Linux-Einführung durch Wolfgang Heuscher:

- Proxy: FBI-WWW.cs.uni-dortmund.de Port 3128
- Links in Taskleiste kreieren: Taskleiste - Spezialknopf - Nicht KDE Programm und dort den Link eintragen



Mindmap.jpg

Abbildung 13.2: FIPs Mindmap

TOPS nächste Sitzung

(siehe nächste Sitzung)

21.04.2005

Abwesend: Jana (krank)

Verspätet: Rene (5 min), Stefan (10 min), Ahmet (55 min)

Sitzungsleitung: Rene Goebels

Protokollführung: Christian Friem

Tagesordnung

1. Begrüßung
2. Formalia
3. Festlegung möglicher Arbeitsgruppen
4. Vortrag Projektmanagement
5. Sonstiges
6. TOPs nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen. Haus Nordhelle hat noch keine Rechnung geschickt. Es fehlen noch ICQ-Nummern. Bitte an Stefan mailen...

Festlegung möglicher Arbeitsgruppen

Es wurden folgende Gruppen gebildet, die sich selbstständig organisieren: Kunden: Niels, Madan, Jana, Rene Informationsbeschaffung: Bertram, Martin, Mehmet, Christian Finanzprodukte: Christoph, Markus, Stefan, Ahmet

Die Gruppen sollen nach einer Woche einen Zwischenbericht ablegen, dann wird die verbleibende Zeit für die Aufgaben festgelegt.

Vortrag Projektmanagement

Die Folien zu Niels Vortrag sind im Folienpaket der Seminarfahrt enthalten.

Sonstiges

Die PG-Kasse wird am Montag gefüllt. Dazu sollte jeder 10 Euro mitbringen.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 28.04.2005

Abwesend: niemand

Verspätet: niemand

Sitzungsleitung: Christoph Hübinger

Protokollführung: Ahmet Kara

Tagesordnung

1. Begrüßung
2. Formalia
3. Endberichte der Kleingruppen
4. Weiteres Vorgehen
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen. Die Abgabe der Seminarausarbeitungen wird um eine Woche verschoben

Endberichte der Kleingruppen

Gruppe Informationsbeschaffung trägt ihr Bericht vor. Diese Gruppe wird sich noch mit der Informationsbeschaffung bezüglich Anleihen beschäftigen.

Gruppe "Finanzprodukte"trägt ihr Bericht vor Es werden im wesentlichen 6 Finanzprodukte vorgestellt. Stefan schlägt vor, den Fokus auf Aktien, Optionsscheine, Anleihen und Fonds zu legen und Rohstoffe und Sparbücher ausser Betracht zu lassen. Als nächstes soll diese Gruppe sich Gedanken darüber machen, mit welcher Datenstruktur die einzelnen Finanzprodukte gespeichert werden könnnen.

Gruppe "Kunden"trägt ihr Bericht vor. Diese Gruppe soll sich Gedanken darüber machen, inwieweit Fuzzy-Logik bei der Modellierung der Kundenpräferenzen zum Einsatz kommen kann.

Weiteres Vorgehen

Bis 9. Mai sollen Kleingruppen fertig werden

In der nächsten Sitzung soll der PG-Fahrplan für das Semester festgelegt werden. Die Minimalziele aus der PG-Beschreibung werden vorgelesen. Sie sollen als Diskussionsgrundlage für den PG-Fahrplan dienen.

Die Tools für Projektmanagement werden zur Wahl gestellt. Die Entscheidung fällt für 'e-groupware'.

Sonstiges

Die Rechnungen von unserer Seminarfahrt sind fertig. Das Geld soll schnellstmöglich an Stefan überwiesen werden.

Es wird entschieden,dass im nächsten Semester eine Seminarphase im Rahmen der PG gemacht wird. Die Seminarthemen sollen im Bereich CI liegen. Das Thema kann jeder selber aussuchen und es mit dem Beträuer besprechen.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 2.5.2005

Abwesend: niemand

Versptet: niemand

Sitzungsleitung: Ahmet

Protokollfhrung: Markus

Tagesordnung

1. Begrung
2. Formalia
3. Zwischenberichte der Kleingruppen
4. Grobe Zielvorgabe
5. Sonstiges
6. TOPS nchste Sitzung

Begrung

Die Sitzungsleitung begrt die Anwesenden.

Formalia

Die Protokollfhrung wird festgelegt. Das Protokoll vom 28.4.2005 wurde angenommen.

Zwischenberichte der Kleingruppen

Die Gruppe Finanzprodukte schlgt vor, Informationen als Graphen zu speichern. Die Gruppe Kunden berichtet, dass sie sich mit der Verfeinerung des Bewertungsschemas befasst und dabei die Einteilung der Finanzprodukte nicht in Risikoklassen sondern in Sicherheitsklassen braucht. Die Gruppe Beschaffung gibt einen kurzen berblick ber ihre Arbeit und empfiehlt die Seite: http://www.finanzprodukte.de

Grobe Zielvorgabe

Martin schlgt zwei Vorgehensweisen zum Entwickeln unseres Programms vor:

1. Ein gutes Modell des Systems erzeugen und dann auf dieser Basis anfangen zu programmieren
2. Button up einen Prototypen mit Grundfunktionen zu entwickeln und diesen dann nach und nach zu erweitern.

Die Gruppe entscheidet sich fr den ersten Vorschlag

Um einen Zeitplan aufstellen zu knnen werden die zu machenden Arbeitsvorgnge erst einmal gesammelt:

- Bewertung von Finanzprodukten
  - Fundamentaldatenanalyse
    - Konjunktureinflussgren
    - Unternehmenskenngren
    - Quellen
    - Indikatoren
    - Persistenz
  - Technische Analyse (wird nur gemacht wenn noch Zeit bleibt)
  - Extraktion aus News
    - Quellen
      - Ad- hoc Meldungen
      - Analysteneinschtzungen
      - Sonstiges
    - Text Mining
      - Filtern von Wrtern
      - Expertenfilter
      - Persistenz
- Entscheidung
  - Empfehlungen fr Kunden, dessen Profil wir kennen
    - Empfehlung von Quellen
    - Kauf- und Verkaufvorschlge machen
  - Regelbasis
    - Logiken
    - Inferenz
    - Revision
    - Persistenz
  - Lernen
    - Adaptive Kundenprofiloptimierung
    - Regelbasis optimieren
  - Ausgabe
  - Vorhandene Methoden
  - Anwenden der Methoden auf unser Problem
- Kunden kategorisieren

- Benutzeroberfche
- Webauftritt
- Datenbank
- Schnittstellen zwischen den Komponenten

Aus Zeitgrnden konnte die Liste nicht vervollstndigt werden und jeder Teilnehmer soll sich bis zur nchsten Sitzung weitere Gedanken dazu machen. Auerdem soll sich jeder eine mglichst realistische Einschtzung des Zeitaufwandes fr die einzelnen Punkte berlegen.

Sonstiges

TOPS nchste Sitzung

(siehe nchste Sitzung)

Sitzungsprotokoll vom 9.5.2005

Abwesend: Ahmet Kara

Verspätet: Jana Ehlers (20 Min)

Sitzungsleitung: Markus Matz

Protokollführung: Niels Pothmann

Tagesordnung

1. Begrüßung
2. Formalia
3. Endberichte
4. zusammentragen der Arbeitsvorgänge
5. Zeitplan erstellen
6. Sonstiges
7. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Top3

Gruppe Kunde

> Fragebogen wurde verändert und vorgestellt Formeln wurden angepasst Punkteverteilung ist abgeschlossen Es gibt 120 Punkte, die mit Faktor 1,2 auf 0..100 Punkte verrechnet werden Anmerkung: Besser wäre es, Felder mit „keine Angabe" zur Verfügung zu stellen, aber hierbei soll nicht allzu viel Zeit verwendet werden.
>
> Kunde wird als Vektor gespeichert, Fuzzy Mengen als Trapeze. Die Formeln für die Kombination von Fuzzy-Mengen wurden vorgestellt.

Gruppe Informationsbeschaffung

Konjunktureinflussgrößen: Indikatoren wurden ermittelt

Nachrichtenbearbeitung:

Problem mit Unternehmensberichten (schlecht zu analysieren / meistens in pdf und erzählender Text).
Format XBRL ist guter Lösungsansatz, wird aber nicht oft verwendet / keine verlässliche Quelle

Problem: Nachrichten werden unstrukturiert dargestellt. Informationen sind schwierig zu extrahieren.
Lösung: RSS Feeds werden von den Anbietern zur Verfügung gestellt, können automatisch heruntergeladen werden mit kostenlosem Tool. Feeds sind in XML Format, geben Kurzbeschreibung der Nachricht.

Problem: RSS Feeds geben Link an, der verfolgt werden muss, um zu den Nachrichten zu gelangen -¿ Problem mit Anpassung der Wrapper.
Lösungsansatz: HTML Seite in Baum umwandeln und anhand der Kurzbeschreibung der Nachricht Zweig im Baum identifizieren und dahingehend abgespeichern.

Weiterer Ansatz: Fragen, ob ein Anbieter uns Nachrichten zur Verfügung stellt, dann müsste man keine große Analyse machen.

Gruppe Finanzprodukte

Abspeicherung von Daten

- XML Dokumente
  - implementieren DTD's, sind erweiterbar. Vorteil: Alte Daten werden nicht verändert.
  - Wenn Struktur zu unübersichtlich wird könnten Links zu anderen XML Dateien verwendet werden.
- Baum
- Datenbank (Postgre SQL)

Top4

- Bewertung von Finanzprodukten
  - Fundamentaldatenanalyse
    - Konjunktureinflussgrößen
    - Unternehmenskenngrößen

        * Quellen
        * Indikatoren
        * Persistenz
    - Technische Analyse (wird nur gemacht wenn noch Zeit bleibt)
    - Extraktion aus News
        * Quellen
        * Textmining
        * Persistenz

- Entscheidung
    - Empfehlungen für Kunden, dessen Profil wir kennen
    - Regelbasis
    - Lernen
    - Ausgabe
    - vorhandene Methoden
    - Anwendung der Methoden auf unser Problem
- Kunden kategorisieren (abgeschlossen)
- Benutzeroberfläche (Webauftritt)
    - Kundenfragebogen
    - Kundenverwaltung
    - Ausgabe
    - Administration
    - Information
    - Layout
    - Persistenz
    - Sicherheit
- Datenbank
    - Technologie
    - Datenmodell
    - Testdaten
    - Interface
    - Persistenz
- Schnittstellen zwischen den Komponenten

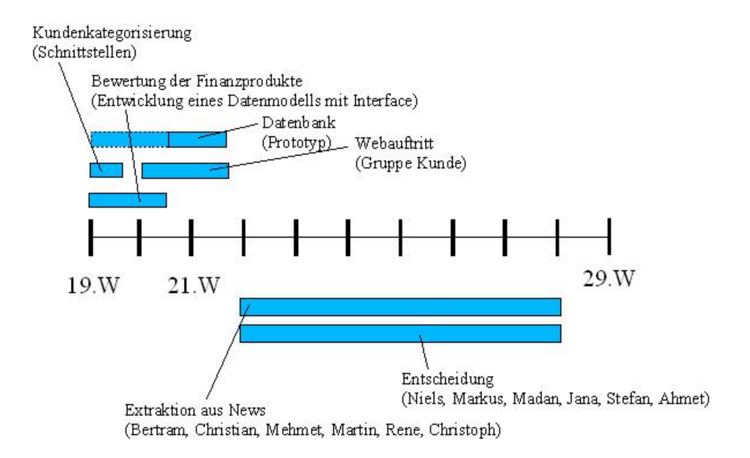


Abbildung 13.3: Zeitplan

Top5

Sonstiges

- IRB Accounts sind freigegeben, werden am Donnerstag verteilt oder können vorher bei Stefan abgeholt werden
- Donnerstag wird der Schrank gefüllt
- Rene ist am 30.5. abwesend
- Kundenfragebogen wurde verteilt

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 12.05.2005

Abwesend: niemand

Verspätet: niemand

Sitzungsleitung: Niels

Protokollführung: Martin

Tagesordnung

1. Begrüßung
2. Formalia
3. Vortrag von Wolfgang
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Vortrag von Wolfgang

Wolfgang war leider nicht da, also den Vortrag (Linux-Einführung) verschoben auf die nächste Sitzung.

Sonstiges

- IRB Accounts wurden verteilt.
- Der Schrank wird nach der Sitzung aufgefüllt.
- Die Kunden-Gruppe hat mit dem UML Tool Umbrello schonmal die
- Kundenschnittstelle definiert.
- Stefan B. meinte, dass man sich schon bald mal Gedanken machen könnte, was unser Programm den leisten soll, also quasi ein Pflichtenheft erstellen.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 19.05.2005

Abwesend: Christian (entschuldigt)

Verspätet: Ahmet (20 Min.)

Sitzungsleitung: Martin

Protokollführung: Stefan R.

Tagesordnung

1. Begrüßung
2. Formalia
3. Endberichte
4. Vortrag von Wolfgang Hunscher
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Endberichte

Die Gruppe "Beschaffung" lieferte einen Bericht zur Bewertung der Finanzprodukte ab.
Inhalt:

- Anwendungsfalldiagramm mit KI / Robot als Aktoren
- Klassendiagramm der einzelnen Finanzprodukte
- Aktivitätsdiagramm eines Wrappers für RSS

Es ist eine Zusammenfassung zur Bewertung von Finanzprodukten im Ordner

`/home/pg473/pg473/Beschaffung`

abgelegt.

Bertram erwähnt die Problematik mit der Datenkonsistenzhaltung.

Vortrag von Wolfgang Hunscher

Leider ist Wolfgang nicht vorbereitet. Daher wird der Vortrag auf die nächste Sitzung verschoben.

Sonstiges

Ein "Prototyp" für den Webauftritt der Anwendung ist erstellt worden und unter

`/home/pg473/pg473/Webentwurf`

verfügbar. (René)

Niels schlägt vor, die eGroupware einzurichten. Jana wird eGroupwarebeauftragte.

Die Gruppe "Beschaffung" wird sich mit der Einschätzung der Kennzahlen von Finanzproduken beschäftigen.

Es wird auf ein Bedienungsdefizit des UML-Tools Umbrello hingewiesen.

Es soll ein Pflichtenheft erstellt werden. Um sich mit dem Konzept eines "Pflichtenheftes" vertraut zu machen, bringt Martin zur nächsten Sitzung ein solches Heft mit, um dann kurz den Aufbau zu erläutern.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll 23.05.2005

Abwesend: niemand

Verspätet: niemand

Sitzungsleitung: Stefan Rosas

Protokollführung: Mehmet Sari

Tagesordnung

1. Begrüßung
2. Formalia
3. Bericht DB
4. Bericht Bewertung
5. Vorstellung eines Pflichtenheftes
6. Vortrag von Wolfgang Hunscher
7. Sonstiges
8. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Bericht DB

- Die Datenbank läuft auf den Servern
- Eine Beschreibung kann man im Ordner /pg473/pg473/JBoss finden

Bericht Bewertung

- Die Gruppe Bewertung der Finanzprodukte hat 9 Kennzahlen ausgewählt, die bewertet werden sollen
- Für die Bewertung der Aktien wird die Fuzzy-Logik herangezogen
- Die Kennzahlen werden Fuzzy-Mengen zugeordnet und durch eine Regelbasis wird die Bewertung festgelegt

Vorstellung eines Pflichtenheftes

- Vorstellung eines Pflichten- und Lastenheftes durch Martin
- Die Unterlagen, die von Martin vorgestellt wurden befinden sich im gemeinsamen Ordner unter Pflichtenheft
- Alle Kleingruppen sollen sich schon Mal Gedanken über Pflichtenhefte für die jeweiligen Gruppen machen

Vortrag von Wolfgang Hunscher

E-Group Ware wurde installiert und wird im Laufe der Woche zum Laufen gebracht

Sonstiges

- Kurzer Bericht der Kundengruppe
- Gruppe Kunde stellt die ersten Layout-Beispiele vor
- Die Gruppen fangen mit der Bearbeitung der 2 großen Theorieblöcke Extraktion aus News und Entscheidung an

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 30.05.2006

Abwesend: René

Verspätet: Ahmet (5 min), Christian, Madan, Stefan (30 min)

Sitzungsleitung: Mehmet

Protokollführung: Bertram

Tagesordnung

1. Begrüßung
2. Formalia
3. Vortrag von Wolfgang
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Vortrag von Wolfgang

Wolfgang war leider nicht da, also wurde der Vortrag auf die nächste Sitzung verschoben.

Sonstiges

- Das Pflichtenheft wurde angefangen, steht im PG Ordner. Jeder soll seinen Bereich dort eintragen, Gedanken über weiters Vorgehen sind erwünscht.
- Egroupware läuft jetzt (auch uniextern)
  URL: http://pg473.cs.uni-dortmund.de/egroupware
  Accounts gibt es bei Jana und Niels
  Wichtig: Nach Benutzung das Ausloggen nicht vergessen!
  Grobe Tasks der Kleingruppen wurde angelegt
- USBSticks können am Rechner ls1pool9 benutzt werden.
  Nach Befehl mount stehen die Daten unter /home/pg473/usbstick

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 02.06.2005

Abwesend: Martin(entschuldigt)

Verspätet: niemand

Sitzungsleitung: Bertram

Protokollführung: Madan

Tagesordnung

1. Begrüßung
2. Formalia
3. 1.Vorstellung der theoretischen Realisierung der fundamentalen Aktienanalyse und der praktischen Probleme
4. 1.Vorstellung des Konzepts der Kleingruppen
5. 1.Vortrag von Wolfgang
6. Sonstiges
7. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Protokoll vom 30.5.05 wurde angenommen.

theoretische Realisierung der fundamentalen Aktienanalyse

Bertram stellt vor, wie die Markteinschätzung über die Fundamentalanalyse berechnet wird und haben es an einem Bespieldatensatz ausprobiert. Weist auf Problem hin, dass Fundamentalanalyse nur eine Tendenz der Aktie angeben kann und Aussage mit Vorsicht zu genießen ist. Kann man sich im Egroupware Ordner unter BewertungAktien anschaun.

Konzept der Kleingruppen

- Extraktion aus News-Gruppe hat ein 7 Schritte System aufgestellt, wie man an Informationen aus RSS-Feeds kommt und damit die Texte bewerten kann. Die Schritte 1-3 hat die Gruppe so weit fertig bearbeitet. Haben eine Email rumgeschickt mit Informationen zu den einzelnen Schritten.

- Die Gruppe Ëntscheidung"führt vor, wie man aus den Eingaben, Kundenvektor und Finanzvektor, mit Hilfe der Fuzzy-Logik zu einem Ranking kommt, in dem die Wertpapiere für den Kunden nach dem höchsten Wert sortiert werden. Mehr dazu steht im Egroupware Ordner Ëntscheidung".

Vortrag von Wolfgang

Vortrag von Wolfgang ist ausgefallen. Wird aufs nächste Mal verschoben.

Sonstiges

- Stefan erinnert daran, dass das Pflichtenheft weiter bearbeitet und in egroupware der Zeitplan fertiggestellt werden soll.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 9. Juni 2005

Abwesend: Jana

Verspätet: Christoph (20 Min.)

Sitzungsleitung: Christian

Protokollführung: René

Tagesordnung

1. Begrüßung
2. Formalia
3. Zwischenberichte der Kleingruppen
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Zwischenberichte der Kleingruppen

Siehe Präsentationsfolien der beiden Teilgruppen im Egroupware

Sonstiges

René hat 5 Euro eingesammelt für die Auffüllung des Schrankes von allen.

Die Sitzung am 13.06. fällt aus, da die Kleingruppen noch Zeit brauchen und sonst nur Begrüssung und Formalia auf der Tagesordnung ständen

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 23.06.2005

Abwesend: Bertram

Verspätet: Jana(10 Min.)

Sitzungsleitung: Christoph Hübinger

Protokollführung: Ahmet Kara

Tagesordnung

1. Begrüßung
2. Formalia
3. Berichte
4. Pflichtenheft
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Berichte

Die Gruppe Extraktion aus Finanznachrichten hat sich mit dem Clustern von Finanznachrichten beschäftigt und hat ihre Ergebnisse präsentiert. Dabei sind sieie folgendermaßen vorgegangen:

- Die Finanznachrichten wurden in XML-Dateien geschrieben. Mit einem Clusteralgorithmus wurde versucht, die Nachrichten zu klassifizieren.
- Mit den Wörtern, die in den Nachrichten vorkommen wurde ein Wörterbuch erstellt.
- Aus der Liste wurden mit Gate Stoppwörter rausgefiltert
- Für jede Nachricht wurde ein Dokumentenvektor mit zwei Spalten erstellt, wobei die Zahl an der ersten Stelle die Stelle des entsprechenden Wortes im Wörterbuchdarstellt und die die Zahl an der zweiten Stelle angibt, wie oft das Wort im Satz vorkommt.

- K-Means-Clusteralgorithmus mit Ähnlichkeitsabstand: Hemmingabstand

Diese Methode wurde auf ca. 3000 Nachrichten ausprobiert. Leider ist das Ergebnis als erfolglos zu bezeichnen, da die Nachrichten, die in das selbe Cluster fielen nichts miteinander zu tun haben.

Es wurde über alternative Lösungsvorschläge diskutiert. Es wurde angemerkt, dass das Problem an dem Wörterbuch liegen könnte. Es wurde vorgeschlagen, ein kleineres Wörterbuch, das von einem Börsenexperten erstellt wurde zu verwenden. Des Weiteren wurde diskutiert, was das Ziel des Clusterns sein soll und es wurde die Frage gestellt, inwiefern die geclusterten Nachrichten behilflich bei der Entscheidung über ein Finanzprodukt sind. Es wurde unter anderem auch vorgeschlagen, zuerst eine Wörterliste zu erstellen und danach diese auf die Nachrichten anzuwenden.

Die Gruppe Entscheidung hat ihre Ergebnisse vorgestellt und eine To-do-Liste angegeben. Die Gruppe geht bei der Erstellung eines Aktienrankings folgendermaßen vor:

- Aus den Angaben im Fragebogenformular wird der Kundenvektor [Sicherheit] [Verfügbarkeit] [Rendite] bestimmt
- Zu allen Aktien werden die Finanzvektoren [Sicherheit] [Verfügbarkeit] [Marketperformer] bestimmt
- Mit Fuzzy Min-Operation wird Kundenvektor mit Finanzvektor kombiniert, erhaltene Flächeninhalte addiert und mit Zugehörigkeitswerten gewichtet
- Größter Flächeninhalt kommt an erste Stelle des Rankings
- Rankings hat größte Übereinstimung mit Kundenprofil

Die Todo-Liste für diese Gruppe sieht folgendermaßen aus:

- Anpassung der Regelgewichte für die verschiedenen Sicherheitsklassen mittels Testdaten
- Berechnen der Branchendurchschnittswerte für die weiteren Branchen
- Verschmelzung von Marketperformer und Rendite (FuzzyMenge kombinieren mit Zugehörigkeitswerten von Marketperformer)
- Einbeziehen von News in die Entscheidung (warten auf Gruppe Extraktion)

Pflichtenheft

Christoph wird das Pflichtenheft für die Gruppe 'Extraktion' übernehmen. Bei der anderen Gruppe ist noch nicht entschieden, wer das Pflichtenheft vorbereitet

Sonstiges

Es wird beschlossen, am Donnerstag, den 30.6. gemeinsam zu grillen. Für den Einkauf wird in der nächsten Sitzung von jedem 10 Euro eingesammelt.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 27.06.2005

Abwesend: Bertram (entschuldigt)

Verspätet: niemand

Sitzungsleitung: Ahmet

Protokollführung: Markus

Tagesordnung

1. Begrüßng
2. Formalia
3. Berichte der Kleingruppen
4. Pflichtenheft
5. Grillen
6. Sonstiges
7. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das letzte Protokoll wird erst heute von Ahmet herumgeschickt.

Berichte der Kleingruppen

- Die Extraktion aus News-Gruppe beschäftigt sich weiterhin mit dem Hammingabstand.
- Die Entscheidungsgruppe ist dabei, die Branchendurchschnitte für jede der zehn Kennzahlen zu berechnen.

Pflichtenheft

In das Pflichtenheft wurde nun auch der Webauftritt aufgenommen und die Gruppe meint, dass sie nun mit dem Pflichtenheft fertig ist. Sie ist allerdings noch offen für konstruktive Kritik.

Grillen

Das Grillen findet am Donnerstag, den 30.6.05 um 17.30 Uhr bei Rene statt. Eine Wegbeschreibung will Rene noch rumschicken. Einkaufsliste für das Grillen:

- Fleisch und Vegetarisches zum Grillen
- Bier / Getränke (Prof. Reusch und Stefan wollen auch jeweils was spendieren)
- Kohle
- Besteck
- Salate
- Brot
- Kruterbutter
- Ketchup

Sonstiges

- Es wurden 10 Euro eingesammelt
- Die Sitzung am Donnerstag wurde auf den Donnerstag Abend verlegt
- Für nächsten Montag soll jede Gruppe ein Resümee ziehen und sich wenn möglich einen Zeitplan für das weitere Vorgehen überlegen, der dann in eGroupware eingetragen werden soll.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 7.4.2005

Abwesend: Bertram (entschuldigt) und Jana (entschuldigt)

Verspätet: Stefan Berlik (10 min.)

Sitzungsleitung: Markus Matz

Protokollführung: Niels Pothmann

Tagesordnung

1. Begrüßung
2. Formalia
3. Berichte der Kleingruppen
4. Pflichtenheft
5. Grillen
6. Sonstiges
7. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Top3

Die Gruppe Extraktion stellt die Ziele und Ergebnisformate vor.
Der Clustering-Algorithmus wurde verbessert indem das Wörterbuch verkleinert wurde. Das neue Verfahren wurde getestet und ausgewertet. Nach wie vor ist es allerdings schwierig, daraus News zu extrahieren (maßgeblich wegen schlechtem Wörterbuch).
Es werden nun alternative Ansätze verfolgt um ein besseres Wörterbuch zu erzeugen. Weiterhin wurden neuronale Netze zur Dokumentenklassifikation vorgestellt. Zu einem bestimmten Kunden soll ein Ranking von Finanz-News erstellt werden.

Die Gruppe Entscheidung stellt vor, was bislang erreicht wurde und was noch erreicht werden soll. Es wurde ein Prototyp implementiert, mit dessen Hilfe die internen Variablen und Gewichtungen zur Entscheidungsfindung angepasst werden sollen. Man erhofft sich davon, die Entscheidungsfindung zwischen

Kunde und Finanzprodukt zu verbessern. Als Ziele wurde insbesondere die Implementation von News genannt, für die jetzt die notwendigen Voraussetzungen vorliegen.

Top4

Stefan hat sich das Pflichtenheft angesehen. Was noch fehlt: Der Kunde müsste einsehen können, warum ihm etwas empfohlen wurde aus dem System.
Kunde müsste dem System Rückmeldungen geben können, damit sich das System anpassen

Top5

Das Grillen wird auf nächsten Donnerstag verschoben (14.7.05)

Sonstiges

Die CI Seminar-Themen und Daten zur Vorbesprechung sind fertig und können auf Stefans Homepage eingesehen werden.

Die Sitzung am Donnerstag fällt aus !!!!
Nächstes Treffen ist dann am Montag den 11.7.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 11.07.2005

Abwesend: niemand

Verspätet: René (15 Min.), Christian (7 Min.)

Sitzungsleitung: Niels

Protokollführung: Martin

Tagesordnung

1. Begrüßung
2. Formalia
3. Berichte der Kleingruppen
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Berichte der Kleingruppen

Gruppe Bewertung:

Implementierung soweit fertig. Versuch noch Marktperformer einzubinden.

Interne Aufteilung der Gruppe:

2 Leute beschäftigen sich mit der DB und 2 Leute beschäftigen sich mit der Anbindung an HTMl bzw. JSP.

Gruppe Extraktion:

Hat jetzt sehr viele Lösungsansätze und muß nun daraus realisierbare auswählen und bewerten.

Sonstiges

- Grillen am Donnerstag den 14.07.2005 um 17:30 bei Rene.
- Die Sitzung am Donnerstag 14.07.2005 fällt dementsprechend aus
- Entory AG: Vorschlag von Markus, dass die Minimalziele im Pflichtenheft ein wenig besser ausgearbeitet werden sollten.
- Anregungen von Stefan: Vielleicht zu Beginn des neuen Semesters eine Zusammenfassung für die Entory AG machen, was wir bis jetzt haben.
- Ziele am Montag den 18.07.2005: Zusammentragen was wir noch tun müssen. Zeitplan erstellen für das nächste Semester.
- Kassensturz: Rene macht KEINEN Kassensturz zum Ende des Semesters
- Betram bezahlt noch 10 Euro für das Grillen
- Stefan R. hatte noch Schwierigkeiten das SDK für J2EE zu instalieren - ¿ Wolfang

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 18.07.2005

Abwesend: niemand

Verspätet: niemand

Sitzungsleitung: Martin

Protokollführung: Stefan R.

Tagesordnung

1. Begrüßung
2. Formalia
3. Berichte
4. Zeitplan
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Berichte

Niels berichtet über die Einarbeitung in die Datenbank. Es wurden dabei Tabellen angelegt und SQL-Befehle ausprobiert. Schwierigkeiten gab es bei der Erstellung einer neuen Datenbank, da die benötigten Benutzerrechte nicht vorhanden waren.

Madan berichtet über die Einarbeitung in Webtechnologien (JavaServerPages, HTML). Eine erste Version der Webpräsenz ist bereits erstellt worden (Entwurf, Realisierbarkeit). Des weiteren ist eine Anbindung an die Datenbank möglich.

Christoph berichtet über den Fortschritt im Extraktionsbereich. Es wurde ein vernünftiges Wörterbuch erstellt (kleiner, handlicher). Der PartOfSpeech-Tagger ist integriert worden.

Zeitplan

Ziele zum Ende der PG

1. Dem Benutzer einen Überblick über die Nachrichten geben:
   - vorsortiert (Suchwörtern)
   - Ranking
2. Entscheidungs- /Bewertungskomponente für Finanznachrichten:
   - vom System
   - durch den Benutzer
3. Fundamentale Bewertungskomponente (vom System)
4. (optional) andere Finanzprodukte (ausser Aktie)
5. Web-Interface: Benutzerführung
6. Dokumentation
7. Test

Aufgaben

zu 1)

- Information Retrieval
- Kategorisierung für die Vorsortierung (Schema, Cluster)
- Ranking bzgl. Schema

zu 2)

- Auf Methoden zur Bewertung / Kategorisierung festlegen (auch: negative Ergebnisse dokumentieren)
- Bewertungsfunktion festlegen für die Benutzer-Bewertung
- Transparent machen der Methoden

zu 3)

- Automatische Aktualisierung der Fundamentaldaten und Durchschnittsberechnung
- Fundamentaldaten in die Datenbank schreiben

zu 5)

- Darstellungsoptionen:

  - Brachensortiert
  - ...
  - Hinweis auf andere Finanzprodukte

- Darstellung:
  - Sicherheit der Entscheidung / Bewertung angeben
  - Sicherheit des Rankings angeben
- (optional) Firmenportrait
- intuitive Benutzerführung
- Kombination von Fundamentaldaten und Extraktion von News

zu 6)

- Benutzerhandbuch
- Entwicklungsdokumentation
- Wartung
- Zwischenbericht und Endbericht
- Testergebnisse

zu 7)

- Modultest
- Testverfahren entwerfen zum Produkttest (effizient!) und anwenden → Ergebnisse dokumentieren
- Validierung → "Feedback Agent"

Zeitplan:
1) und 3): parallel in alten Gruppen
2) neue Gruppen bilden / Methoden testen, implementieren

Sonstiges

Nächste Sitzung ist Mitte Oktober.
Die Sitzung am Do, 21.7.05, fällt aus.

TOPS nächste Sitzung

(siehe nächste Sitzung)

13.1.2 Protokolle 2. Semester

Sitzungsprotokoll vom 20.10.2005

Abwesend: Niels(entschuldigt)

Verspätet: niemand

Sitzungsleitung: Mehmet

Protokollführung: Madan

Tagesordnung

1. Begrüßung
2. Formalia
3. Besprechung der Modellierung + Zeitplan
4. Gedanken darüber, was am 27.10 präsentiert wird
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Besprechung der Modellierung + Zeitplan

- Ahmet stellt von der Gruppe Entscheidung einen Gesamtsystemüberblick aus Sicht seiner Gruppe vor und gibt Details über die Komponenten DB-Controller, DB-Schema und Web-Logik preis. Zu dem werden zwei Sequenzdiagramme zu den Vorgängen, LogIn und Empfehlung berechnen, vorgestellt

- Bertram erläutert von der Gruppe Extraktion aus News deren Architektur von dem Zusammenspiel der Komponenten (z.B. Wrapper, DB, BenutzerInterfaces, Logik) im Gesamtsystem Newsberechnung wird erst einmal so implementiert, dass der Nutzer des Systems eine News bewerten (Relevanz, Tendenz) soll (Easy-IR) News-Bewertung vom System schwierig, aber wird versucht

- Beide Gruppen stellten deren noch zu erledigenden Aufgaben vor und zeigten Ansätze von einem Zeitplan
- Zeitplan wird am Montag in einem Treff der Kleingruppen fertig gestellt und am Dienstag präsentiert

Gedanken darüber, was am 27.10 präsentiert wird

- Montag wird eine Prototyp-Präsentation erstellt, die dann ebenfalls am Dienstag vorgestellt wird, dabei wird dann auch der Überblick des Gesamtsystemes beider Gruppen zusammengestellt

Sonstiges

- Schrank im PG-Pool wird wieder aufgefüllt
- Radio wird von Rene wieder mitgebracht

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 27.10.2005

Abwesend: Jana (krank)

Verspätet: Markus (15 min), Christoph (30 min)

Sitzungsleitung: Bertram Bödeker

Protokollführung: Christian Friem

Tagesordnung

1. Begrüßung
2. Formalia
3. Präsentation
4. Vorstellung des Zeitplans mit Aufgabenverteilung
5. TOPs nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen. Datenbank und Tomcat sind eingerichtet, eine Passwortänderung wurde nicht beschlossen.

Präsentation

Ahmet, Martin und René haben einen Probedurchlauf für den Vortrag gemacht. Alle waren mit dem Ergebnis zufrieden.

Vorstellung des Zeitplans mit Aufgabenverteilung

Die Aufgabenverteilung wurde besprochen. Die Zuordnung ist auf dem Zettel Zwischenziele um das System zum Laufen zu bringen zu sehen. Als Zeitraum wurde zunächst eine Woche (bis zum 3. Nov. 05) veranschlagt.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 3. November 2005

Abwesend: Jana (entschuldigt)

Verspätet: keiner

Sitzungsleitung: Christian

Protokollführung: René

Tagesordnung

1. Begrüßung
2. Formalia
3. Kurzer Überblick der Kleingruppen/Zwischenergebnisse
4. Zeitplan
5. Klassendiagramm
6. Sonstiges
7. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll vom 27.10.05 wird angenommen, mit der Änderung, dass Task 4b von Martin und Ahmet anstelle von Bertram und Ahmet bearbeitet wird.

Kurzer Überblick der Kleingruppen/Zwischenergebnisse

- Bertram stell den DB Controller vor. Genaueres ist in der Präsentation nachzulesen, die sich im Egroupware System befindet
- Die Kleingruppen sollen sich überlegen, welche Daten sie von der DB benötigen, damit die Funktionalitäten programmiert werden können, ohne dass später wieder allzuviel geändert werden muss
- Dokumentationen sollen im Egroupware System im Ordner Dokumentation gespeichert werden
- Madan und Stefan sind noch mit dem Fundamentaldatenwrapper beschäftigt
- Seitenstruktur der Webpage steht, Teilgruppe lernt noch HTML

- Corpora Speicherung in der Datenbank wurde von Christoph verbessert
- Aktienempfehlungen, die über das Ranking hinausgehen, bleiben weiterhin eine optionale Funktionalität des Programms

Zeitplan

Jede Teilgruppe aktualisiert die Dauer und den Bearbeitungsstatus seiner Task im Egroupware System

Klassendiagramm

Ï Ahmet und Bertram erstellen ein Klassendiagramm für FIPs, welches als Diskussionsgrundlage dient

Sonstiges

Ï PG Treffen bleiben weiterhin dienstags und donnerstags um 8:15, jedoch wird bei Nichtbedarf auch mal ein Termin ausfallen

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 3. Januar 2006

Abwesend: Mehmet (entschuldigt)

Verspätet: Christian, Ahmet, Stefan, Niels (5 Min.), Markus (10 Min.)

Sitzungsleitung: René Goebels

Protokollführung: Jana Ehlers

Tagesordnung

1. Begrüßung
2. Formalia
3. Zwischenergebnisse der Kleingruppen
4. Zeitplan
5. Klassendiagramm
6. Sonstiges
7. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Zwischenergebnisse der Kleingruppen

- Kleingruppe Website fast fertig, aktuelle Version in egroupware noch zu ändern:
- Schieberegeler Risikofreudigkeit fehlt
- Wdh. Email unnötig
- Passwort als neue Textbox aufnehmen
- bei 'mein Portfolio' sollen die News mit in der Tabelle zur Aktie verlinkt sein
- in der Menüleiste links muss 'mein Portfolio' ergänzt werden
- die Syntax zu 'unsere Bewertung' soll bei + und bleiben, Zahlenwerte des Programms werden normalisiert und in diese Symbole umgewandelt

- die Seiten 'Empfehlungen' und 'News' sind noch nicht fertig
- das Impressum muss noch ergänzt werden
- Kleingruppe DatenbankController
- Datenbankschema ist noch nicht ganz fertig, daher auch der DBController noch nicht
- setzen sich mit Bertram und Christoph zusammen
- einige DB-Verbindungen und -Methoden laufen; es fehlen noch Ein- und Ausgabedaten
- an manchen Stellen ist unklar, was dem System zu welchem Zeitpunkt bekannt ist
- Kleingruppe Fundamentaldaten-Wrapper
- holt Daten von Onvista und schreibt sie in DB
- Branchendurchschnitte werden per Hand errechnet und eingefügt
- Kleingruppe Bewertung
- 'Portfolio' ist rausgefallen
- Berechnung der Fragebogendaten ist implementiert
- die News-Tendenz ist in die Bewertung einbezogen
- es fehlt noch die Methode für das Herausziehen der Gesamttendenz aus DB
- das Format der News ist noch nicht bekannt
- Kleingruppe Datenbank
- das Verknüpfen der Unternehmen läuft
- muss noch mit DB verbunden werden

Zeitplan

- 1. A und G wird ergänzt 1. B-F ist fertig 2. A wird angepasst 2. B ist gestrichen 2. C ist fertig 3. A-B wird ergänzt 4. A wird ergänzt (einige Methoden) 4. B und D fehlt noch 4. C kann angefangen werden, wenn Website fertig ist 4. E-G ist in Bearbeitung
- bis Donnerstag erledigen die Kleingruppen die notwendigen Ergänzungen
- ab Donnerstag wird das System dann zusammengebaut

- parallel dazu beginnen wir, uns in die Themen 'Klassifikationsmöglichkeiten' (6 B) und 'Lernen' (6 E) einzuarbeiten. Kleingruppen dazu: SOMS (neuronale Netze) Bertram, Martin ART-Netzte Christoph, Madan Entscheidungsbäume René, Ahmet Konzeptlernen Niels, Christian Suche nach anderen Klassifikationsmöglichkeiten Stefan, Mehmet Lernmöglichkeiten Jana, Markus

Klassendiagramm

Klassendiagramm und DB-Tabellen wurden z.T ergänzt und müssen noch weiter ergänzt werden. Die Klasse Fragebogen wird von Niels ergänzt. Die Klasse Task-Manager kann erst später entstehen.

Sonstiges

Keiner weiß genau, wofür das 'S' in FIPs steht.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 15.11.2005

Abwesend: niemand

Verspätet: niemand

Sitzungsleitung: Christph Hübinger

Protokollführung: Ahmet Kara

Tagesordnung

1. Begrüßung
2. Formalia
3. Erste Rückmeldung von der Zusammenfügung der Systemkomponenten
4. Klassifizierungsmethoden
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Erste Rückmeldung von der Zusammenfügung der Systemkomponenten

Martin hat die bis dahin fertig gestellten Komponenten zusammengefügt. Bestimmte Funktionalitäten, wie Berechnung und Ausgabe eines Rankings funktionieren. Einige Funktionalitäten, die mit Gate gelöst werden sollen, fehlen. Außerdem gibt es noch Mängel bei der Berechnung des Kundenprofils über den Fragebogen. Die aktuelle Version des Systems liegt im CVS. Als weiteres stehen die Tests an.

Stefan Berlik schlägt vor, dass die Tests systematisch und dokumentiert gemacht werden sollen, damit die gewünschten Funktionalitäten garantiert werden und die Dokumentation direkt in die Enddokumentation eingebunden werden kann. Niels und Christian werden die Tests durchführen und dokumentieren.

Klassifizierungsmethoden

Stefan berichtet, dass die Firma Entory der Kleingruppe, die die beste Lösung für das Problem der Extraktion aus News findet mit einem kleinen Preis auszeichnen will.

Die Gruppe, die sich mit Konzeptlernen beschäftigt hat, stellt fest, dass diese Methode doch nicht geeignet zu sein scheint für unser Problem.

Die Gruppe, die sich mit anderen Lernmethoden auseinandersetzt hat, hat sich mit den Support-Vektor-Maschinen beschäftigt.

Die Gruppe Lernen hat sich überlegt, dass das System verfolgen könnte, für welche Branchen sich der Kunde mehr interessiert, um die Vorschläge dementsprechend anzupassen.

Stefan Berlik macht den Vorschlag, dass man nun auch nach Lösungen suchen sollte, die die Semantik von Texten bearbeiten, um wenigstens einfache Sätze interpretieren zu können.

Renè und Ahmet werden sich mit der semantischen Analyse von Texten beschäftigen.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 17.11.2005

Abwesend: Ahmet, Jana(entschuldigt)

Verspätet: niemand

Sitzungsleitung: Niels

Protokollführung: Martin

Tagesordnung

1. Begrüßung
2. Formalia
3. Berichte der Kleingruppen
4. Berichte über die Tests
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll vom 15.11.2005 wurde noch nicht bestätigt.

Berichte der Kleingruppen

Kleingruppe Konzeptlernen:

Nils stellt vor, warum das mit dem Konzeptlernen nicht funktioniert.Gründe dafür sind:

- Konzepte müssen bekannt sein
- Für jedes Konzept müssen geeignete Attribute in jedem Text vohanden sein
- Für jedes Konzept müssen positive und negative Beispiele vorhanden sein

Kleingruppe ART:

Madan stellt ART-Netze vor und begründet warum ART2a Netze jetzt näher untersucht werden (unüberwachtes Lernen, leicht zu implementieren, konvergiert

schnell). Chritoph erklärt einen ART2a Algorithmus (Parameterinitialisierung, Initialisierung der Klassenmatrix, Preprocessing, Klassifizierung und Trainingsphase)

Kleingruppe SVM:

Stefan stellt SVM vor (überwachte Lernmethode) zur Lösung linear separierbarer Kategorisierungsprobleme. Definition von Hyperebenen. Begründungen warum SVM gut zur Textklassifikation geeignet ist (gute Generalisierung, schnelle Verarbeitung von vielen Attributen) Mehmet gibt Ausblicke und Asätze, wie auch nicht lineare Kategorisierungsprobleme mit SVM gelöst werden können.

Kleingruppe SOM:

Bertram definiert SOM und zeigt an einem Beispiel wie sich ein SOM organisiert. Dabei gibt Stefan B. Tips, wie SOMs auch mit großen Vektoren und einer festen Gitterstruktur arbeiten (Bsp. Klassifizierung von Audiosignalen) Bertram gibt einen Ausblick zum Thema, wie sich eine bewertete Nachricht zum DAX verhält. Er hat die Implementierung begonnen, dass zu einer Nachricht die Aktienkurse in drei Zeitintervallen (nach 1 Tage,3 Tage und einer Woche) aus dem Internet geladen werden und analysiert wird, wie sich der Kurs verhält und wie die Nachricht zur Aktie bewertet worden ist. Bertram möchte die Implementierung bis zur nächsten Sitzung abgeschlossen haben. Damit andere Gruppen schonmal mit den Ergebnissen arbeiten können die Definition der Schnittstelle wie folgt:

```
interface testvektor {
int nr;
String titel;
String text;
Timestamp datum;
int bewertung;
String aktie}
```

Dabei ist die Ausprägung der Bewertung:

- 0=keine
- 1=—
- 2=–
- 3=-
- 4=0
- 5=+
- 6=++

- 7=+++

Kleingruppe Lernen:

Gibt noch keine vorstellbaren Ergebnisse.

Kleingruppe Semantik:

Gibt noch keine vorstellbaren Ergebnisse.

Berichte über die Tests

Nils berichtet, dass die Tests noch schwierig sind, da die Daten sehr miteinander verknüpft und die Ergebnisse schwer nachzuvollziehen sind. Die Funktionalitäten der Webseite müssen noch getestet werden.

Madan macht den Vorschlag/Erstellt eine todo Datei im CVS, in der alle noch zu erledigenden Implementierungs/Verbesserungsaufgaben stehen. Jeder soll diese Datei ansehen und Aufgaben davon erledigen und kurz kennzeichnen, was er gemacht hat oder ob eine Aufgabe gerade in Bearbeitung ist.

Sonstiges

Entory (Markus) möchte gerne die zuletzt geänderten Dokumente sehen. Die sind aber schwierig im egroupware zu finden. Deshalb wird jetzt ein change.log im egroupware eingeführt, dass zu pflegen ist und in dem alle Änderungen oder Neuerungen von Dokumenten aufgeführt sind.

Der Zeitplan soll bis/am Dienstag aktualisiert werden.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 22.11.2005

Abwesend: Stefan B.(entschuldigt)

Verspätet: Christian, Stefan R., Jana (5 min)

Sitzungsleitung: Ahmet

Protokollführung: Markus

Tagesordnung

1. Begrüßung
2. Formalia
3. Zwischenergebnisse der Kleingruppen
4. Zeitplan
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das letze Protokoll wurde angenommen.

Zwischenergebnisse der Kleingruppen

- Bertram hat eine neue DB erstellt in der es neue Einträge zur Bewertung von Aktien anhand ihrer Kurse gibt
- SOMS: Die Implementierung ist fertig und steht ist Modul im CVS verfügbar. Die Dokumentation ist im egroupware zu finden.
- ART- Netze: die Implementierung ist fast fertig und wird bis Dienstag ganz fertig sein.
- Support- Vektor Maschinen: Es gibt Probleme, da die gleichen Wörter in den beiden Klassen vorkommen können, die Gruppe wird sich aber noch weiter damit befassen und schauen, ob und wie das Problem behoben werden kann.
- Lernen: Benutzerstatus wird eingeführt, der angibt wie gut der Benutzer die Nachrichten in der Vergangenheit eingeschätzt hat. Je besser die Vorhersagen des Kunden waren, desto stärker fliesen seine neuen Einschätzungen in die Berechnung des Rankings der Aktien mit ein.

- Semantik: Eine alte Theorie wurde gefunden die die Semantik mit Hilfe der Prädikatenlogik abbildet soll. Jedoch gibt es keine Umsetzung und die Theorie wurde auch nicht mehr weiter entwickelt, da die Syntaxanalysen sinnvoller erschienen. Daher wird es wohl schwer werden eine eigene Semantikanalyse zu erstellen!
- Testen: Einige Fehler sind aufgefallen und korrigiert. Weitere Fehler sind noch in der TODO- Liste im CVS und sollen von allen Mitgliedern der PG behoben werden.
- Das zuordnen von ISIN zu den Nachrichten funktioniert nun auch.

Zeitplan

Bis Dienstag arbeiten die Kleingruppen an ihre Projekten weiter und präsentieren dann ihre Ergebnisse. Jede Gruppe dokumentiert ihre Ergebnisse und stellt diese ins egroupware.

Sonstiges

- Es wurde beschlossen, dass die Sitzung am Donnertag den 24.11.05 ausfällt.
- In der nächsten Sitzung am Dienstag, den 29.11.05 werden wieder 10 Euro von jedem für den PG-Schrank eingesammelt.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 29.11.2005

Abwesend: Christoph, Niels(entschuldigt)

Verspätet: niemand

Sitzungsleitung: Markus

Protokollführung: Stefan R.

Tagesordnung

1. Begrüßung
2. Formalia
3. Ergebnisse der Kleingruppen
4. Zeitplan
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Ergebnisse der Kleingruppen

Gruppe "Lernen" (Bericht von Jana):

Die Benutzerbewertungen der Nachrichten sollen bei der Berechnung eines Ranking mit einbezogen werden. Dabei sollen zeitnahe Tendenzen und Relevanzen stärker in die Berechnung einfließen als ältere Bewertungen. Des weiteren wird eine benutzerabhängige Gewichtung der Bewertungen verwendet. Dazu ist ein Bonussystem entwickelt worden, welches bei zutreffenden Bewertungen eines Benutzers, den neueren Bewertungen dieses Benutzers ein grösseres Gewicht bei der Berechnung des Ranking zuweist. Die Bewertungen inkompetenter Benutzer werden (aufgrund der falschen Bewertungen) abgeschwächt.

Gruppe "Entscheidungsbäume" (Bericht von Rene und Ahmet):

Es wurde eine allgemeine Einführung gegeben (Definition, Algorithmen zur Erstellung von Entscheidungsbäumen, Repräsentation der Dokumente). Im Verlauf

des Preprocessing wurden die vorhandenen Texte intelektuell ausgewertet, um wichtige Attribute zu extrahieren (Feature Selection). Die Klassifikation erfolgt mit Hilfe von Vergleichsdokumenten/-vektoren. Die Problematik besteht darin, eine geeignete Attributkombination zu finden. Die ersten Testläufe lieferten - trotz der heuristischen Vorgehensweise - überraschend gute Ergebnisse.

Die erzeugten Bäume sind zu statisch. Ein neuer Ansatz besteht darin, auf der Basis von Kombinationen von Wörtern aus einem Finanzwörterbuch, eine Menge von Klassifikationsfunktionen zu erzeugen und diese zu evaluieren. Zu diesem Zweck sind bereits korrekt klassifizierte Nachrichten notwendig (solch eine Trainingsmenge ist auch für andere Kleingruppen eine Voraussetzung). Für die Erzeugung der Funktionen wurde ein Algorithmus vorgestellt. Von den Anwesenden wird die hohe Komplexität dieses Ansatzes angemerkt.

Gruppe "Konzeptlernen" (Bericht von Christian):

In der Einführung wurden die Begriffe "Konzept", "Konzeptlernen" und "Versionenraum" erläutert. Konzeptlernen ist ein überwachtes Lernverfahren (benötigt also positive und negative Beispiele). Die theoretischen Überlegungen über die (Nicht-)Anwendbarkeit des Verfahrens führten zu dem Versuch eines praktischen Tests.
Für der Implementation des Versionenraumlernverfahrens wurden Tests mit unterschiedlich grossen Wörterbüchern durchgeführt. Die Ergebnisse sind niederschmetternd. Bei zu kleinen Wörterbüchern verlor der Versionenraum seine Konsistenz. Zu grosse Wörterbücher stießen an die Grenzen der Java-Virtual-Machine (out of memory). Es konnte kein korrektes Konzept gefunden werden, womit der Bereich des Konzeptlernens für die Problemstellung der PG ausgeschlossen werden kann.

Gruppe "ART2a-Netze" (Bericht von Madan):

Der Algorithmus für das unüberwachte Lernen wurde implementiert. Um die Effizienz zu steigern, muss noch ein spezielles Preprocessing durchgeführt werden. Dieses besteht darin, gewisse Teilmengen eines Finanzwörterbuches durch Oberbegriffe zusammenzufassen ({ Gewinn,Verlust, Rückgang } -¿ Unternehmensentwicklung).
Die Möglichkeit der Klassifikation ist nicht vorhanden (jedenfalls nicht beim derzeitigen Stand der Dinge). Allerdings können zu einer Nachricht ähnliche Nachrichten zurückgegeben werden. Dies wäre für Benutzer ein interessantes Feature (ähnlich zur kollaborativen Filterung beim Online-Kaufhaus "Amazon").

Gruppe "SOM" (Bericht von Martin):

Es wurde eine Aufgliederung in 2 Bereiche durchgeführt. In Bereich des unüberwachten Lernens existiert eine fertige Version eines SOM-Verfahrens. Im Bereich des

überwachten Lernens wurde die Methode der STM (semantische Topic-Maps) vorgestellt. Diese Methode sieht ein umfangreiches Preprocessing der vorhandenen Daten vor. Es müssen z.B. Stopwörter entfernt, Sätze gesplittet/vereinfacht und sogenannter Spam gelöscht werden. Des weiteren ist eine Änderung der Repräsentation der Nachrichten nötig. Die ursprünglichen Dokumentenvektoren (bag-of-words Darstellung) müssen in (Unternehmen, Graph)-Tupel umgewandelt werden. Als Graph verwendet man dabei sogenannte Topic-Maps. In Kombination mit einem Finanzthesaurus werden die Abhängigkeiten der einzelnen Attribute in den Topic-Maps dargestellt.
Für die Anwesenden scheint dieser Ansatz die grösste Erfolgswahrscheinlichkeit bzgl. der Analyse von Texten zu besitzen.
Aus Zeitgründen muss Bertram's Bericht über multi-layer-Netze verschoben werden.

Gruppe "SVM" (Bericht von Mehmet):

Die Support-Vektor-Maschinen benötigen bereits klassifizierte Beispiele, um ein Klassifikationsmodell zu finden. Dazu müsste man in unserem Fall eine Menge von Nachrichten lesen und entsprechend eines Klassenmodells klassifizieren. Aus diesen Beispielen können dann Dokumentvektoren erstellt werden, die dann als Eingabe für die SVMs dienen. Von einer eigenen Implementierung einer SVM wurde abgeraten ("Implementierung kann tükisch werden"). Daher wird die Lernumgebung Yale ("Yet Another Learning Environment" vom Lehrstuhl 8) verwendet, die bereits mehrere SVM-Implementationen beinhaltet.

Zeitplan

Die Gruppen Entscheidungsbäume, ART2a, SOM und SVM sind noch mit ihren Themen beschäftigt.
Die Gruppen Lernen und Konzeptlernen sollen sich mit der Todo-Liste auseinandersetzen.

Sonstiges

Der Kassierer (Rene) hat von allen Anwesenden jeweils 10 Euro für die Befüllung des Schrankes eingenommen. (ausstehend Niels, Christoph, Stefan B.)

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 12.1.2005

Abwesend: Christian Friem (entschuldigt)

Verspätet: niemand

Sitzungsleitung: Martin Prause

Protokollführung: Niels Pothmann

Tagesordnung

1. Begrüßung
2. Formalia
3. Berichte der Kleingruppen
4. Zeitplan
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Top3

Bertram stellt Multilayer-FF-Netze vor und Ansätze zur Umsetzung in Bezug auf das Problem der Nachrichtenklassifizierung, insbesondere den Aspekt der Codierung und Darstellung der Nachrichten-Texte. Es wurde über diverse Ansätze der Codierung und Durchführung diskutiert. Wenn sich herausstellen sollte, dass man die Nachrichten nicht nach dem Performer klassifizieren kann, wird erhofft, wenigstens die Relevanz einer Nachricht bestimmen zu können. Es kam allgemein das Bestreben auf, möglichst bald eine Festlegung für das weitere Vorgehen der Klassifizierung von News zu realisieren.
Folgende Konzepte stehen uns dafür zur Auswahl:

- SVM (Mehmet + Stefan)
- ART 2a (zurückgestellt)
- Entscheidungsbäume (Ahmet + René → Vertiefung (besseres Preprocessing))

- Konzeptlernen (fällt raus, da nicht mächtig genug)
- ML-Feed-Forward (Bertram → Vertiefung)
- SOM (Martin → Vertiefung (besseres Preproc.))

weitere Aufgaben:

- 'gutes' Wörterbuch bestimmen (Jana + Markus)
- Spam-Filter (Madan, Christoph)
- Testen (Niels + Christian)

Top4

Dienstag fällt die Sitzung aus!
Nächsten Donnerstag sollten erste Ergebnisse der Kleingruppen feststehen!

Top5

Wer noch Pfandflaschen aus dem Pool hat soll sie bitte wieder abgeben.

Die Rechner im Pool haben kein funktionierendes Openoffice mehr, bzw. ein Rechner fehlt! Bertram sagt Wolfgang bescheid oder löst das Problem selbst.

Anmerkung für nächsten Donnerstag: Es kam die Idee auf, mit der PG über den Weihnachtsmarkt zu gehen. Es soll in der nächsten Sitzung darüber abgestimmt bzw. ein Termin bestimmt werden.

Sonstiges

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 08.12.2006

Abwesend: Martin (entschuldigt)

Verspätet: Christian (10 Min.), Madan (10 Min.)

Sitzungsleitung: Stefan Rosas

Protokollführung: Mehmet Sari

Tagesordnung

1. Begrüßung
2. Formalia
3. Ergebnisse der Kleingruppen
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Ergebnisse der Kleingruppen

Gruppe to do's Niels berichtet:

- to-do-Liste wurde erweitert
- einige Punkte wurden schon bearbeitet und einige Punkte sind noch dazu gekommen
- weil ein wichtiges Interface fehlt, wollen sich Nils und Christian mit der Gruppe Web-Entwicklung zusammensetzen

Gruppe Wörterbuch Markus berichtet:

- graphische Schnittstelle für Nachrichten wurde erstellt
- es müssen nur noch die einzelnen Wörter durch markieren in das Wörterbuch eingefügt werden
- bis Dienstag soll das Wörterbuch fertig sein

Gruppe Spamfilter Christoph berichtet:

- für das Problem wurden Support Vektor Maschinen benutzt
- hierzu wurden 700 Nachrichten klassifiziert -¿ es ist jedoch kein vernünftiges Ergebnis dabei raus gekommen
- als zweiter Ansatz wird jetzt ein bayesscher Filter benutzt -¿ dieser muss jetzt mit den Nachrichten trainiert werden

Gruppe Entscheidungsbäume Ahmet berichtet:

- existiert schon eine Implementierung von Entscheidungsbäumen in WEKA -¿ diese soll in die Implementierung der Gruppe Entscheidungsbäume integriert werden
- die Dokumentation zu Entscheidungsbäumen ist schon fertig gestellt

Gruppe Support Vektor Maschinen Stefan berichtet:

- es wurden knapp 200 Nachrichten klassifiziert (per Hand)
- es gibt eine Fehlerrate von knapp 27 Prozent, die nicht reduziert werden kann
- die Idee, die Textklassifikation mit SVM's zu lösen wird verworfen -¿ es soll eine Dokumentation über das Vorgehen mit SVM's geschrieben werden

Gruppe SOM Bertram berichtet:

- es wurde schon bereits das meiste implementiert
- es wurden auch schon mehrere Tests durchgeführt
- es wurde herausgefunden, dass die Zeit für das Trainieren der neuronalen Netze akzeptabel ist
- dieser Ansatz wird weiter verfolgt

-¿ bis nächste Woche sollen die Kleingruppen an ihren Aufgaben weiterarbeiten

Sonstiges

Am nächsten Montag ist der Besuch des Weihnachtsmarktes geplant. Für alle, die Interesse haben: Treffen ist am Montag um 18.30 an der Reinoldikirche (an der Pylone).

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 13.12.2005

Abwesend: Jana(entschuldigt)

Verspätet: Christian(10 min)

Sitzungsleitung: Mehmet

Protokollführung: Madan

Tagesordnung

1. Begrüßung
2. Formalia
3. Ergebnisse der Kleingruppen
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Ergebnisse der Kleingruppen

Gruppe: todo´s

Berichtet von: Niels

- Interface für Bewertungsabgaben ist fertig gestellt
- User kann Bewertung zu News abgeben
- weitere Punkte der ToDo-Liste sind eher Schönheitskorrekturen, die allerdings nicht die Funktionalität beeinträchtigen (Martin kümmert sich um den Punkt, dass nur eine Verbindung zur DB von Nöten ist)
- Testen bis Donnerstag das Gesamtsystem

Gruppe: Wörterbuch

Berichtet von: Markus

- graphische Schnittstelle ist endgültig fertiggestellt

- jeder soll bis Donnerstag einige Nachrichten durchlesen und für Aktien relevante Wörter ins Wörterbuch schreiben

- zu dem sind Synonyme und Antonyme zu bestimmen

Gruppe: Spamfilter

Berichtet von: Madan

- der bayessche Filter klassifiziert Junks mit Hilfe von Blacklists (eine für "gute"Wörter und eine für ßchlechte"Wörter) mit 80% richtig

- das Ergebnis kann wahrscheinlich nur noch leicht verbessert werden

Gruppe: Entscheidungsbäume

Berichtet von: Ahmet

- das Programm von WEKA steht im CVS und muss jetzt noch getestet werdens

Gruppe: SVM

Berichtet von: Mehmet

- Dokumentation wird bis Donnerstag fertig gestellt

Gruppe: SOM

Berichtet von: Bertram

- FF-Netze müssen mit relevanten Daten getestet werden
- erweiterte SOMs (semantische topic maps,...) müssen im nächsten Jahr von allen bearbeitet werden

Stand der Klassifikation

- SVM, Art2a, Konzeptlernen gescheitert
- FF-Netze, E-Bäume noch zu testen
- erweitertes SOMs noch zu implementieren

Allgemeine noch zu erledigende Aufgaben:

- Wörterbuch erstellen
- Multilayer-FF-Netze testen

- Entscheidungsbäume testen
- Struktur der Gesamtdokumentation wird von Christian und Rene bis morgen rumgeschickt, so dass sich die anderen Kleingruppen Gedanken zu den noch fehlenden Inhalten machen können

- Dokumentation der Kleingruppen zur Klassifikation
- Spamfilter muss noch eingebaut werden

Sonstiges

Nix

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 15.12.2005

Abwesend: Jana (entschuldigt)

Verspätet: niemand

Sitzungsleitung: Madan

Protokollführung: Bertram

Tagesordnung

1. Begrüßung
2. Formalia
3. Ergebnisse der Kleingruppen
4. Zeitplan
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Ergebnisse der Kleingruppen

- Mehmet über SVMs:
  die Dokumentation ist fertig und wird ins Egroupware gestellt
- Ahmet über Entscheidungsbäume:
  Wollte testen, es fehlten Testdaten, Mehmet hat welche, diese werden von ihm ins egroupware gestellt, Bertram gibt den Hinweis, dass in der DB testdaten unter nachricht klassifizierte Testnachrichten (abhängig vom Kursverlauf) vorhanden sind.
- Niels über TODO-Liste:
  viele Fehler sind Behoben, neue wurden erkannt und behoben, es bleibt ein kleiner Schnitzer in der Bewertungsfunktion von Aktien und ein großer Schnitzer beim Setzen vom Kundenstatus. Bearbeitung läuft...
- Zusätzliche Punkt auf der TODO Liste:

1. Nur Nachrichten zu bewerteten Aktien werden angezeigt $\leadsto$ Methode im DBcontrol ändern auch systembewertete Nachrichten anzeigen
2. Junk-Filter in Wrapper einbauen $\leadsto$ anstoßen von testdatencreator, doppelte/leere Nachrichten killen, Einzeiler löschen
3. Wörterbuch mit Synonymen, evtl. Topic Map
4. Statements weiter in eigene Datei kopieren
5. Vereinfachung beim Einfügen von Aktien
6. Quellen für Wrapper prüfen

- Bertram über SOMs und MLFFNN:
  Das Neuronale Netz kann nun getestet werden: Eingabe: Wörterbuchdatei, Datei mit klassifizierten und vorverarbeiteten Nachrichten (genaueres cvs Projekt NeuronalesNetz). Ist aber noch nicht geschehen da noch kein sinnvolles Wörterbuch vorhanden

- Struktur des Endberichtes
  es wurde sich auf eine vorläufige Struktur des Endberichtes geeinigt. Sie steht im Latex Format unter /Endbericht/Endbericht-v0.8.tex im egroupware. Diese wird von Madan und Niels weiter bearbeitet um detailiertere Punkt des Entscheidungsprozesses zu ergänzen.

Zeitplan, Aufgabenverteilung

0	1	2	3	4	5	6	7
←Junk→	(Christoph, Madan)						
←	TODO	→ (Niels)					
←	SOM, STM, FFN			→ (Martin, Bertram)			
←	Entscheidungsbäume			→ (Ahmet)			
←	Doku Kleingruppen						→ (alle)
←	Gesamt Dokumentation						→ (Mehment, René, Christian, Jana)

Sonstiges

nichts

TOPS nächste Sitzung

Die nächste Sitzung ist am Dienstag, dem 2. Januar 2006
(TOPs siehe nächste Sitzung)

Sitzungsprotokoll vom 3. Januar 2006

Abwesend: Christian (entschuldigt)

Verspätet: Markus (15 Min.)

Sitzungsleitung: Bertram Bödeker

Protokollführung: Jana Ehlers

Tagesordnung

1. Begrüßung
2. Formalia
3. Berichte der Kleingruppen
4. Sonstiges
5. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen. Bericht: Gestern/heute konnte sich keiner im Rechnerpool oder bei egroupware einloggen.

Berichte der Kleingruppen

- Spamfilter (Madan, Christoph): Spamfilter ist fertig eingebunden.
- Endbericht (Christian, Mehmet, René, Jana): Vorläufige Endbericht-Struktur steht. *to do: Endbericht-Struktur im CVS angelegen; Chapter-Ordner durchnummerieren, Section-Ordner nicht (damit noch verschiebbar).*
- Entscheidungsbäume (Ahmet): Entscheidungsbäume wurden 3x getestet auf Wörterbuch von Markus & Jana, von je 38 Nachrichten wurden 27/19/21 unbekannte Nachrichten richtig klassifiziert.
- Testen und Reparieren (Niels, Stefan, Madan): Schieberegler sind eingebaut. News zum Unternehmen lassen sich jetzt per Anklicken anzeigen; leider noch einige irrelevante Nachrichten dabei. Performance des DB-Controllers wurde verbessert. Nachrichten lassen sich blättern. *to do: Suchen, wo die 0,0-Werte herkommen. Prüfen, wieso GATE irrelevante Nachrichten ausgibt. Bertrams Testvektor, der die Aktienkurse verfolgt, in Kundenbewertung einbinden (Jana).*

- SOMS (Martin): Implementierung begonnen. *to do: fertig stellen*
- neuronale Netze (Bertram): Implementierung begonnen. *to do: fertig stellen*

Sonstiges

Zeitplan: Bis Donnerstag to dos weitermachen und an Dokus schreiben. Ahmet meldet sich für Mitte März ab.

TOPS nächste Sitzung

(siehe nächste Sitzung)

12.01.2006

Abwesend: niemand

Verspätet: Ahmet (5 min), Christoph (15 min)

Sitzungsleitung: René Goebels

Protokollführung: Christian Friem

Tagesordnung

1. Begrüßung
2. Formalia
3. Bericht neuronale Netze / SOMs
4. Berichte der anderen Kleingruppen
5. Zeitplan
6. Sonstiges
7. TOPs nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen. Berichtigung zum letzten Protokoll: Jana und Madan machen die Struktur der Endpräsentation.

Bericht neuronale Netze / SOMs

Bei den ML-FF Netzen war das Training laut Bertram erfolgreich, die Testläufe brachten jedoch keine guten Ergebnisse. Auch bei den SOMs gab es keine verwertbaren Ausgaben. Martin und Bertram werden ihre Ergebnisse dokumentieren.

Berichte der anderen Kleingruppen

Endpräsentation: Es sollen noch folgende Tops hinzugefügt werden: Preprocessing (und die Schwierigkeiten dabei), ca. 1 Folie für jedes Klassifizierungsverfahren incl. der Probleme, am Ende sollte der Ausblick noch geändert werden. Enddokumentation: Jeder schreibt seine eigenen Protokolle in Tex. René hat dafür die Vorlage gemacht. To do: Kapitel 1: René, Mehmet, Christian Kapitel 2: René, Mehmet, Christian Kapitel 3: René, Mehmet, Christian Kapitel

6: Bertram, Christoph Kapitel 9: HTML: Madan, Stefan; Wörterbuch: Jana; RSS: Martin; Testdaten: Bertram Kapitel 11: René, Mehmet, Christian Kapitel 12: René, Mehmet, Christian Kapitel 13: René, Mehmet, Christian Kapitel 14: René, Mehmet, Christian

Zeitplan

Bis zum 19. Januar sollten die einzelnen Teile des Endberichtes (möglichst) fertig gestellt sein, am 24. Januar (Dienstag) soll dann alles durchgesprochen werden.

Sonstiges

Nächste Sitzung ist am 19.01.2006.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Sitzungsprotokoll vom 19.01.2006

Abwesend: Rene

Verspätet: niemand

Sitzungsleitung: Christian Friem

Protokollführung: Ahmet Kara

Tagesordnung

1. Begrüßung
2. Formalia
3. Enddokumentation
4. Endpräsentation
5. Sonstiges
6. TOPS nächste Sitzung

Begrüßung

Die Sitzungsleitung begrüßt die Anwesenden.

Formalia

Die Protokollführung wird festgelegt. Das Protokoll der letzen Sitzung wird angenommen.

Enddokumentation

Es haben noch nicht alle die Dokumentation ihrer Teilbereiche beendet. Bis Donnerstag, den 26.01. soll die Dokumentation jedoch komplett fertig werden. Es wird vorgeschlagen, dass sich alle die Gesamtdokumentation durchlesen, damit Fehler berichtigt und Wiederholungen in der Dokumentation herausgestrichen werden können.

Endpräsentation

Madan stellt die aktuellen Folien für die Endpräsentation vor. Bisher sind 33 Folien erstellt; die Zahl wir jedoch wahrscheinlich auf 40 steigen.

Die Folie zu Preprocessing muss überarbeitet werden.

Eine beispielhafte Nutzung des Systems soll mit einem Video, in dem einige Grundfunktionalitäten ausgeführt werden präsentiert werden. Martin schlägt vor, dass man für die Aufnahme die freie Software WINK nutzen soll.

Jede Kleingruppe, die sich mit einer Klassifikationsmethode beschäftigt hat, soll eine Folie für die entsprechende Methode vorbereiten und sie bis Montag an Madan oder Jana schicken.

In die Folie Ausblick sollen folgende Inhalte ein: Weiterentwicklung des Preprocessing Neben Kaufvorschlägen soll das System auch Verkaufsvorschläge anbieten Die Semantische Analyse der Textanalyse soll weiterentwickelt werden

Sonstiges

Die Sitzung am Dienstag, den 24.01. fällt aus.

TOPS nächste Sitzung

(siehe nächste Sitzung)

Literaturverzeichnis

[Cau] Jörg Caumanns. A fast and simple stemming algorithm for german words. ftp://ftp.inf.fu-berlin.de/pub/reports/tr-b-99-16.ps.gz.

[Fou] The Apache Software Foundation. Lucene - eine open-source suchmaschine. HTTP://lucene.apache.org.

[IfmS] Universität Stuttgart Institut für maschinelle Sprachverarbeitung. Treetagger - ein sprachunabhängiger wortart-tagger. HTTP://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger-de.html.

[Pai96] C.D. Paice. Method for evaluation of stemming algorithms based on error counting. *Journal of the American Society for Information Science. 47(8)*, pages 632–649, August 1996.

[Scha] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. HTTP://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger2.pdf.

[Schb] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. HTTP://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf.